

**PSYCHOLOGICAL
TESTS AND
PERSONNEL
DECISIONS**

151.2
CRO

CRONBACH AND GLESER

7637/2

8 3.50



Bureau Ednl. Res. Research
Date 16.2.59
No. 1327

151.2
CRO







Psychological Tests and Personnel Decisions



**PSYCHOLOGICAL
TESTS AND
PERSONNEL
DECISIONS**

LEE J. CRONBACH AND
GOLDINE C. GLESER

University of Illinois Press, Urbana, 1957

151.2
CRO

FOREWORD

The program of research reported here was supported by the Office of Naval Research under contract N6ori-07146. Additional support was received from the Bureau of Educational Research, University of Illinois.

The studies were initiated by Lee J. Cronbach. Goldine C. Gleser was an informal consultant from the earliest stages, and has shared responsibility for the later phases of the work and for this report. Eugene I. Burdock was research associate on the project during the year 1952-1953. Research assistants who have contributed to the project are Jack C. Merwin, Peter Erve, and Jean Denby Manis. Vera Stroud has prepared the manuscript for the report, and Henry Plahn prepared the drawings. The counsel of Robert Vogeles of the Art Department of the University of Illinois Press is gratefully acknowledged.

LEE J. CRONBACH
University of Illinois College of Education

GOLDINE C. GLESER
University of Cincinnati School of Medicine

CONTENTS

List of Figures	x
List of Appendices	xii
I. The Tester and Decision Theory	1
Origins of Decision Theory	2
Character of This Report	3
II. Types of Personnel Decisions	6
Institutional vs. Individual Decisions	6
Maximization as a Decision Principle	8
Assumption of a cardinal utility scale	8
Assumption of <u>a priori</u> knowledge	9
Two-person games	9
Classification, Placement, and Selection	10
Constraints upon decisions	11
Sequential and Single-Stage Strategies	12
A Taxonomy of Decision Problems	13
III. Characteristics of Decision Problems	16
Strategies	17
Evaluation of the Decision-Making Procedure	19
Outcomes	19
Prediction of outcomes	20
Valuation of outcomes	20
Expected utility from a strategy	21
Payoff Functions Relating Utility to Aptitude and Treatment	22
Multivariate response surfaces	24
Assumption of first-degree payoff function	26

IV.	Selection Decisions with Single-Stage Testing	27
	The Interpretation of Validity Coefficients: Earlier Work . .	27
	Basis for Judging Gain from Testing	30
	<u>A priori</u> population	30
	Best a <u>priori</u> strategy	31
	Contribution of various types of tests	31
	Fixed-Treatment Selection	32
	Linear relation of utility to validity	33
	Selection with Adaptive Treatment	37
	A general payoff surface	37
	Utility in adaptive selection	39
	Comparison with benefit from fixed treatment.	40
	Implications for choice of tests	43
	The Interpretation of Validity Coefficients in Selection . . .	44
	The linear relation	44
	The Taylor-Russell results	45
	Parabolic relations	46
V.	Two-Stage Sequential Selection	48
	Possible Uses of Sequential Methods	48
	Sequential acceptance testing.	48
	Other sequential decisions	49
	Determination of Optimum Strategy	50
	Alternative testing plans	50
	Evaluation assumptions	52
	Location of cutoffs	54
	Comparison of sequential and non-sequential strategies .	55
	Effect of variation in cost	58
	Effect of variation in r_1	58
	Pre-Reject and Pre-Accept Strategies	59
	Amount of Testing Saved.	60
VI.	Placement Decisions	64
	Relation of Utility to Validity	65
	Fixed-treatment conditions.	65
	Validation of tests for placement with fixed treatments . .	67
	Adaptive-treatment conditions	69
	Effect of increasing fineness of discrimination	71
	The Interpretation of Validity Coefficients in Placement and	
	Measurement.	74
VII.	Efficient Testing Procedures	78
	Adjustment of Test Length	78
	Optimum length for a single test.	78
	Optimum battery length.	80

	Distribution of effort in two-stage sequential testing . . .	81
	Multiple-Stage Testing	82
	Sequential testing for placement purposes.	85
VIII.	The Bandwidth-Fidelity Dilemma	88
	Utility of a Test in Compound Decisions	89
	Distribution of Effort in a Multi-Score Battery.	90
IX.	Classification Decisions	98
	Classification with Fixed Treatments.	99
	Classification with Adaptive Treatment.	104
	Design of Testing Procedures.	106
	Single-stage batteries	106
	Sequential testing.	107
X.	Evaluation of Outcomes.	109
	Institutional Evaluation of Outcomes	111
	Comprehensiveness of evaluations.	111
	Combination of outcomes by empirical analysis.	112
	Inferring values from decisions	114
	Explicit value judgments.	115
	Evaluation for Individual Decisions.	116
XI.	The Importance of Decision Theory	119
	Implicit Assumptions in Test Theory.	119
	Decision Theory Compared to Alternative Models	121
	The Value of General Ability Tests	124
	The Value of Wide-Band Procedures	128
	A Final Word	132
	Appendices.	134
	Bibliography.	162

LIST OF FIGURES

1.	A classification procedure	10
2.	Placement and selection procedures	11
3.	Schematic view of a decision process	16
4.	Three payoff functions	22
5.	Relation of gain in utility to validity (r_{ye}) and selection ratio in fixed-treatment selection	34
6.	Utility as a function of selection ratio and cost	35
7.	Average gain in utility per person accepted	36
8.	Expected payoff as a function of aptitude and treatment	39
9.	Comparison of gains from fixed and adaptive selection	41
0.	Relation of gain in utility to validity (r_{ys}) and selection ratio in adaptive-treatment selection	43
1.	Sequential selection procedure	51
2.	Alternative methods of selection	52
3.	Gain in utility with sequential and non-sequential strategies. . .	55
4.	Gain in utility with sequential and non-sequential strategies. . .	56
5.	Gain in utility with sequential and non-sequential strategies. . .	56
6.	Comparison of sequential and best non-sequential strategy . . .	57
7.	Utility as a function of cost of second test	58
8.	Utility as a function of relative validity of tests	59
9.	Utility with Pre-accept and Pre-reject strategies	59
0.	Comparison of Pre-reject to Sequential and best non-sequential strategy	60
1.	Cutting scores on first test when $Y^1 = 0$	61
2.	Cutting scores on first test when $Y^1 = 1$	62
3.	Proportion of persons given second test	63
4.	Relation of gain in utility to validity (r_{ys}) and selection ratio in adaptive-treatment placement	69

25.	Optimum cutting scores for adaptive placement	70
26.	Increase in benefit from placement with increased number of treatments	72
27.	Possible payoff function for a single treatment	75
28.	Utility as a function of test length	79
29.	Optimum length of test as a function of other parameters	80
30.	Utility in two-stage selection with varying distribution of effort.	82
31.	Utility in multi-stage selection as a function of selection ratio .	84
32.	Utility in multi-stage selection as a function of cost	84
33.	Utility from comparable non-sequential and multi-stage procedures.	85
34.	Relation of utility to bandwidth	93
35.	Quality of men assigned with unidimensional and differential information	102
A-1.	Chart for determining optimal cutting scores on second test . .	149

LIST OF APPENDICES

1. Utility in fixed-treatment selection and placement	134
2. A payoff surface depending on a univariate aptitude	139
3. Utility with adaptive treatment	141
4. Two-stage sequential strategy	145
5. Optimum length of a test	150
6. Strategy for multi-stage selection	154
7. Distribution of effort in compound decisions	157

THE TESTER AND DECISION THEORY

Our society continually confronts people with decisions for which they have inadequate information. It is for this reason that psychological and educational tests exist. Some of the problems on which tests are brought to bear are purely individual: the uncertainties of a boy trying to choose a career, or of a young couple trying to decide whether they are suited for marriage. Equally numerous are the occasions when an administrator, teacher, or clinician turns to tests for assistance in making decisions about many people. The personnel manager wishes to know whom to hire; the military psychologist determines which men are adequately trained and ready for duty; the teacher inquires whether his class should be taught at a rapid or a slow pace. There is no end to such examples of the role of tests in decision making.

It is therefore desirable that a theory of test construction and use consider how tests can best serve in making decisions. Little of present test theory, however, takes this view. Instead, the test is conceived as a measuring instrument, and test theory is directed primarily toward the study of accuracy of measurement on a continuous scale. Hull in 1928 voiced a principle that has been the root of nearly all work on test theory: "The ultimate purpose of using aptitude tests is to estimate or forecast aptitudes from test scores" (43, p. 268). It is this view which we propose to abandon. We acknowledge the usefulness of accurate estimation -- but we maintain that the ultimate purpose of any personnel testing is to arrive at qualitative decisions such as those illustrated in the preceding paragraph.

The value of a test depends on many qualities in addition to its accuracy. Especially to be considered are the relevance of the measurement to the particular decision being made, and the loss resulting from an erroneous decision. Recommendations regarding the design, selection, and interpretation of a test must take into account the characteristics of the decisions for which the test will be used, since the test which is maximally effective for one decision will not necessarily be most effective elsewhere.

An appropriate test theory can evolve from a general and systematic examination of the decision problems for which tests are used, and of the demands these problems place upon the test. In such a study of decision making and its implications for testing, we are fortunately able to draw on extensive mathematical contributions, many of them recent.

The following comment by a leading contributor to decision theory, M. A. Girshick, indicates the need for study of this sort and the difficulties:

Statistical decision theory was originated some 15 years ago by Abraham Wald. In the past decade, great strides have been made in the field by Wald and others. However, so far these developments have had but little impact on experimental research in the social and physical sciences. There appear to be two basic reasons for this: one is the natural lag between theory and practice which so often occurs in science; the other, which in the present case may be more fundamental, is that decision theory to date has been too much concerned with the mathematical foundations of the subject and less with its immediate application. Curiously enough, here is a situation in which the foundational development, difficult as it is, is easier than the application to actual problems. . . .

. . . Decision theory is in the process of development and is not yet a completed science. It takes for its domain the problem of rational behavior in the face of unknown states of nature. It gives a logical foundation and framework to mathematical statistics which it has never had before. It grapples with the problem of design of experiments in a manner never before attempted in statistics. It insists that cost considerations and consequences of decisions be taken into account in every statistical investigation. In doing so, it has brought sequential theory into the general framework and has made sequential decision procedures the rule rather than the exception. It bridges the gap existing in classical statistics between testing hypotheses and estimation by showing that the distinction lies mainly in [the collection of alternative actions that may be taken] . . . , which in one case is finite and in the other infinite. In its attempt to clarify the nature of the statistical decision process, it exposes the limitations as to what is attainable in the face of ignorance, and raises serious problems concerning rational behavior. . . . Finally, because decision theory deals with the problem of decision making in its greatest generality, every particular statistical problem can be immediately placed in a general framework and thus exhibit its ramifications and implications. (36, p. 448, 464)

ORIGINS OF DECISION THEORY

Despite the recency of its development, decision theory is closely related to long-standing problems in the social sciences, and is being used to some extent by psychologists, economists, and sociologists (see 65). Within psychology alone, the theory has been brought to bear on problems ranging from learning experiments to studies of attitude measurement.

Economists have been concerned with decision processes for a long time. Businesses are continually faced with choices among alternative courses of action. Principles are needed which will assist managers in making the most profitable or most beneficial choices among products, among investments, etc. Serious limitations have been found in the theories of classical economics which emphasize the welfare of the individual entrepreneur. Most conspicuously, they fail to consider how decisions are reached by coalitions of individuals with different interests. This criticism led von Neumann and Morgenstern to study how an individual should make decisions when he is one of a group of participants in a game or a market. The resulting "Theory of Games" (67) proved to have great interest not only for economists, but also for military planners.

The work of Wald, who extended the statistical theory of testing hypotheses into a general "statistical decision theory" (69), paralleled and has merged with game theory. Wald's interest was partly instigated by problems of inspection and quality control in industrial production, where decisions to accept or reject manufactured articles are required. (Such inspection has an obvious similarity to one variety of personnel testing.) Both the statistician and the game theorist are concerned with decisions in the face of an uncertain future, but in game theory the uncertainty comes from the prospective selfish actions of competitors, while in statistical decisions the uncertainty comes from random variation in events.

Studies of decisions take into account the risk, benefit, or "utility" of various courses of action. Much thought has been given to problems of defining or estimating such utilities. Economists and mathematicians have tried, for example, to estimate what hypothetical utilities, attached to various consequences, could account for a person's decision to gamble even though he knows that the "house percentage" prevents his winning over the long term. Experimental psychologists have recently begun to study how people actually choose between courses of action, that is, what utilities are consistent with their actions (65).

Even where it is possible to determine what benefit each individual will receive from various courses of action, the best decision is hard to select if the interests of many individuals are at stake as, for example, in economic decisions of government (4). One might assess the total effect of a decision by combining the benefits received by all individuals, but this can scarcely be done unless we can measure all persons' utilities on the same scale. Is the benefit rendered by reducing a rich man's taxes by \$100 equal to the benefit when a poor man's taxes are reduced by \$100? A whole body of theory referred to as "welfare economics" has tried to resolve dilemmas encountered in balancing the welfare of many persons affected by a decision.

A tremendous amount of knowledge has been developed around the general topic of decision problems. Since the tester is concerned with decision making, it is reasonable to expect significant understandings to result from restating testing problems in such a way that this knowledge can be brought to bear.

CHARACTER OF THIS REPORT

Our study did not begin as an attempt to translate utility and decision theories into psychometric terms. Rather, it began with recognition of certain questions about tests which we could not answer adequately under then-existing formulations. One question, for instance, was how one might evaluate a test battery which measures several aptitude dimensions. Hitherto, testers have discussed only the validity of each single score or of the combined test against any single criterion. Except for Horst's very recent work (40), no

effort has been made to consider the total contribution of the test in making decisions over all criteria. In trying to formulate theoretical models which would permit study of such questions, we were led inevitably to examine the utility of personnel decisions. It was at this point that we recognized our approach as a case within Wald's more general decision model. This comprehensive model, we found, clarified a wide variety of testing problems.

We have not drawn all possible inferences from decision theory. Indeed, we barely touch on some issues which patently lead to deeper and deeper inquiries. Furthermore, decision theory has been growing at a rate which has quite prevented us from exploiting the significance for the tester of new developments. There is no reason to think that this flow of new materials will abate. The majority of the recent publications belong to the higher stratospheres of mathematics, and are to be comprehended only dimly by non-mathematicians. We have confined our own study to the simplest and most definite mathematical methods. No doubt some person with greater mathematical competence than ours can extract from this literature much further material for the tester.

The conceptual tools of decision theory make the tester aware of problems which have hitherto been minimized or overlooked. We believe that our re-statement of testing problems in these new terms will assist all users of tests to understand what assumptions they are making and how adequate their procedures are.

At a more technical level, this report examines specific principles of test development and use. Some of the generalizations are restatements of results already in the psychometric literature, which take on new importance, new interrelations, and new meaning in the decision framework. Others are new developments entirely. We shall not list the contents of ensuing chapters here. The reader may obtain an idea of what is to be covered by examining the bibliography at the end of the book; the papers cited (excepting the mathematical sources) indicate the range of psychometric topics with which we deal. Our analysis will extend or reinterpret these papers, or integrate them with other parts of test theory. Since the topics include efficient design of tests, construction of selection batteries, interpretation of validity coefficients, and use of tests for individual assessment, we anticipate that the results will concern a large and diverse audience of test users.

The reasoning of this report is mathematical rather than empirical. A mathematical treatment has both advantages and hazards. The advantages lie in the precision with which conclusions can be stated, the finality with which they can be established, and the wide range of circumstances to which a derivation can apply. In contrast, an empirical study covers only a few particular circumstances and obtains results which are perturbed by various sorts of sampling error.

The disadvantage of the mathematical attack is that it involves assumptions which may not adequately describe real conditions. At times it is even necessary to make assumptions about postulated variables which have never

been observed. In our model, for example, it is assumed that the contribution of a person to an institution, and the costs of testing, can be evaluated in some tangible, countable unit. Girshick's comments on this difficulty are worth quoting:

Here again we see that decision theory demands a great deal of the decision maker. It demands that he be in a position to evaluate numerically for every possible state of nature in the situation under consideration the consequences of any of the actions he might take. It has been argued by many that no human being possesses the ability so to evaluate the utility of the various actions in all possible states. . . . The inability of the decision maker to formulate clearly the loss function is, in fact, a stumbling block in determining what a rational mode of behavior is for him. More bluntly, it is impossible to tell a person what is an optimal way for him to behave if he is unable to formulate clearly what he is after. . . . Decision theory acts as a gadfly to the research worker. It says to him: You cannot solve your problem unless you more clearly define your goal and the consequences of your decisions. Such a prodding is likely to be healthy. (36, p. 463)

Possibly a major contribution of the approach through decision theory is that it points clearly toward a variety of needed empirical studies. Thus our argument indicates that many present personnel procedures are based on assumptions regarding the interaction of the characteristics of the individual and the nature of the treatment to which he is assigned. These implicit, widely employed assumptions have been tested sketchily if at all. A similar finding is that tests used for placement of individuals (among various levels of instruction, for example) should be validated in ways quite different from those in general use.

To provide the general reader with an overall grasp of concepts and results, the main body of this report contains a minimum of technical detail and mathematical reasoning. The detailed mathematical argument is found in a series of technical appendices.

This report does not attempt to present decision theory per se. A mathematical presentation of decision theory may be found in Theory of Games and Statistical Decisions, by Blackwell and Girshick (6). The latter author has presented an elementary survey of basic concepts in the field (36), and a popularized treatment is offered in Bross' Design for Decision (15). The subject of utility analysis has been reviewed by Adams (1) and Edwards (31), the latter review being particularly directed toward psychologists.

Decision theory is provocative, and forces one to alter his accustomed thought patterns. We anticipate that each reader will find a different facet of our argument important for him, and often his thoughts will veer off into pathways not covered by this investigation. The intended contribution of this monograph is, in a word, to stir up the reader's thoughts.

2

TYPES OF PERSONNEL DECISIONS

Any situation where a person is confronted with alternative courses of action is a decision problem. A test theory should encompass in some manner all the various decisions for which tests are used. The reference to "personnel testing" in our title may suggest that we are concerned only with the decisions of industrial and military personnel management, but this is not the case. Our discussion embraces problems of guidance, clinical assessment, and teaching, as well as the problems of personnel management in the narrow sense.

Each type of decision makes somewhat different demands upon tests, and this has important implications for test theory. Indeed, it could be said that test theory must be developed differently for each type of decision. For theoretical purposes, however, it is not appropriate to separate guidance from clinical problems, or problems of teaching from those of industry. Instead, it is necessary to develop rubrics which classify problems according to the formal characteristics significant for decision theory.

INSTITUTIONAL VS. INDIVIDUAL DECISIONS

Institutional decisions such as selection of employees may be distinguished from individual decisions such as choice of a vocation. In the typical "institutional" decision, a single person makes a large number of comparable decisions. The "individual" decision is one in which the choice confronting the decision maker will rarely or never recur.

Examples of institutional decisions are classification of military recruits, screening of pupils to identify who should be studied by counselors, and determination of appropriate therapy for psychotic patients. In any of these instances, decisions are made regarding many people, using a constant philosophy or value system. The hospital wishes to use its therapeutic resources most efficiently; the industry wishes to choose employees who will in the long run contribute most to its balance sheet. One seeks the decision or policy which will yield as much benefit as possible to the institution. Individual interests may

be taken into account, but only insofar as they affect the realization of the goals of the institution. From the viewpoint of an admissions officer, the best use of space in a medical school is to admit those who have the greatest probability of success, considering both motivation and ability. Between two candidates, the institutional value system would require impartial choice of the one with the better prognosis.

In an individual decision, the best course of action depends on the individual's value system and varies from one individual to another. A particular goal which would be worth any risk to one individual may have little value to another. Thus one boy applying for admission to medical school might value highly even a small chance of success, and would regard a shift from medicine into teaching or pharmacy as abandoning all his aspirations; another having comparable prospects of success might be contented or even relieved to abandon a medical career. Probability of success cannot be the sole consideration in the individual decision.

In the institutional decision, we may think of the decision maker as trying to maximize the benefit from a whole series of similar decisions. That is, he seeks a policy which will work best "on the average" over many decisions about admission or job assignment or therapy. Since each decision involves the same set of values, he can combine different decisions and strike some type of statistical balance which gives the best overall outcome.

The individual decision is often unique. The choice may occur only once in a lifetime. Even where the decision can be "remade" at a later time if the first course of action works out badly, the original decision has an uncancellable influence on the welfare of the individual. A poor choice of curriculum at the outset of a student's college education will continue to handicap him long after he has discovered his error and changed to a more suitable curriculum. It is meaningless to speak of "averaging his risk", since he makes only a few such decisions at most.

The insurance business is founded on just this distinction between institutional (or collective) decisions and individual decisions. The company can pool its decisions; it charges a premium such that the company's loss over all risks underwritten will be balanced by the totalled premiums. The premium is further increased by margins for safety, for administrative expense, and for profit. The purchaser is involved in only a few possible insurance transactions. In each, he decides to insure or not insure -- i.e., to incur a small and certain loss, or to risk a large loss at a low level of probability. If he insures, his premium is larger than his pro rata share of the risk. A man considering a large number of insurance transactions would find this course of action uneconomical because he loses on the average. He might more profitably make other financial provisions for covering possible losses. The man considering only a few insurance policies cannot expect that the outcomes for him will come close to the statistical average. He must evaluate the discrete alternative risks according to his personal scale of values.

Test theory, as it now stands, is relevant chiefly to institutional decisions. Regression formulas, for example, are designed to minimize the average squared error of estimate. Institutional decisions lend themselves to "strong" mathematical and statistical treatment, making generalizations possible. For this reason, our report deals almost entirely with institutional decisions. Many important personnel problems may require a decision theory based on the individual's value system. Counseling, for example, has moved more and more in the direction of encouraging individuals to make their own decisions. Some aspects of individual decisions will be discussed briefly in Chapter X.

MAXIMIZATION AS A DECISION PRINCIPLE

Mathematical decision theory is highly general, encompassing many possible principles for arriving at a decision. In some situations a conservative principle which avoids any large loss is preferable; the purchaser of an insurance policy is following such a strategy. In other situations, choices are preferred which offer even a small probability of obtaining a large gain; many individuals seem to adopt this principle when gambling. In institutional statistical decisions, the most generally useful strategy is one which maximizes the average gain (or minimizes average loss) over many similar decisions. We shall employ this principle in developing test theory regarding institutional decisions. Even with this restriction, many important lines of attack are opened, indeed far more than we can exploit at present.

Assumption of a cardinal utility scale

In order to define an expected gain and maximize it, it is necessary to make an important assumption about the utility scale on which the outcomes of possible decisions are evaluated. We must assume that the value of various outcomes can be expressed in "equal units of satisfaction" or the like, which are additive over many decisions. Where students are being selected for medical school, for example, an evaluation in terms of the average grade implies that admitting a group who will attain 30% A, 40% B, and 30% C grades is equivalent in value to admitting B students only. It is evident that such an assumption will not always be acceptable. One decision maker might prefer the former policy on the basis that it admits some highly superior students who will bring credit to the university. Another decision maker might prefer the second policy because the narrow range of ability makes the group easier to teach. The assumption of a cardinal scale of utilities is highly restrictive. Where it is applicable, it leads to relatively simple generalizations about best policies.

Much decision theory is based on the weaker assumption that outcomes may be evaluated on an ordinal scale. This requires simply that the decision maker be willing to order the expected outcomes in terms of his preference. Such ordinal utilities may not be averaged over decisions. Ordinal utilities, and

problems associated with them, are under active study by many persons at present. One reason for this interest, apart from the purely mathematical attractiveness of the problems, is that ordinal utilities have been found a superior basis for economic theory (1). In psychology, Coombs (22) has pointed out that the investigator who restricts himself to cardinal assumptions is unable to attack certain problems adequately. These include probability learning, level of aspiration, trouble-shooting, preference among rewards or among wagers, etc. Ordinal models are being used in some current research on these topics. Out of this work may come practical methods for dealing with ordinal utilities, and important insights for test theory. In our opinion, however, such developments will bear primarily on individual decisions rather than institutional decisions.

Assumption of a priori knowledge

An investigator who wishes to maximize his expectations must have a firm basis for these expectations. An insurance company can successfully predict the net outcome of its wagers solely because of its accumulated experience with the distribution of losses for individuals grouped on the basis of age, occupation, state of health, etc. Knowledge of the expected distribution of outcomes (with and without information from testing) is likewise necessary in personnel decisions. Approximate distributions are ordinarily obtained by administering the test to a sizeable sample of individuals, and using the distributions thus found as a basis for future policy. It is necessary to assume, of course, that the relevant features of the situation remain stable.

Two-person games

Much of the effort in decision theory and the theory of games has been to avoid assumptions such as those made above. von Neumann and Morgenstern (67) point out that many decision problems take the form of a "zero-sum two-person game", in which each person tries to increase his payoff at the expense of the other. The player cannot depend on statistical experience to predict his opponent's move. When one player knows what move the other is going to make, he may be able to take advantage of this knowledge. A course of action which could have great value "if all goes well" is not a sound choice if it leaves a serious weakness for the opponent to capitalize on.

In such a case, it has been suggested that instead of trying to maximize expected payoff it might be better to follow a "minimax" principle. The minimax principle is to select always that course of action which will yield the least loss when one's opponent makes the least favorable response. The minimax principle is highly conservative, embodying the view that "if anything can go wrong, it will". Wald (69) has pointed out that statistical decisions can be described as two-person games, where the decision maker is playing against Nature. The decision maker tries to guess (from partial information) what conditions exist in the particular situation that confronts him; Nature has chosen

one from a wide range of alternatives. The minimax principle can be applied here also, but so pessimistic a view is doubtfully appropriate in statistical decisions, for Nature is presumably indifferent rather than antagonistic. Many attempts have been made to propose other decision principles (65), all of which aim to avoid assuming knowledge of an a priori distribution.

No present alternative appears to be a better basis for making institutional decisions by means of tests than the assumption of a known joint distribution of test score and outcome. The hazards of this assumption should be borne in mind. Work in game theory may someday suggest a superior attack upon testing problems.

CLASSIFICATION, PLACEMENT, AND SELECTION

Each decision in personnel work is an attempt to determine what should be done with one or more individuals. Sometimes the question is whom to hire and whom to reject. Sometimes the question is for what job a given recruit or applicant should be trained. Sometimes the decision is among educational or therapeutic methods. Letting the word treatment take a sufficiently broad meaning to cover all such examples as these, we may say that every personnel decision involves assigning each individual to an appropriate treatment. The number of individuals under consideration and the number of treatments available in any particular decision problem may be large or small.

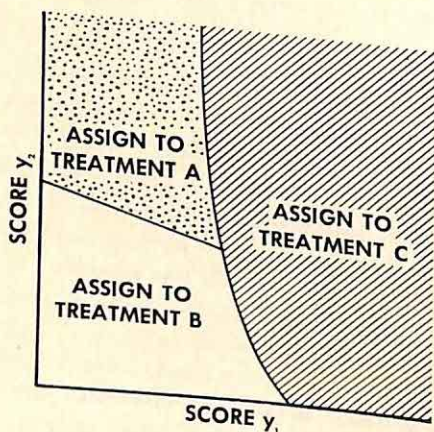


Figure 1. A classification procedure

In its most general form the decision problem is one of classification. Classification is thus understood to include guidance decisions as well as assignment to industrial jobs. It includes psychiatric diagnosis and similar

discriminant problems involving no quotas, as well as classification to fill quotas. Information is used as an aid to classification; this information may be continuous or qualitative, and may involve one or several dimensions. Figure 1 illustrates schematically the general nature of the classification problem.

It is helpful to consider separately two special cases of classification which are much simpler to analyze than the general problem. The first case

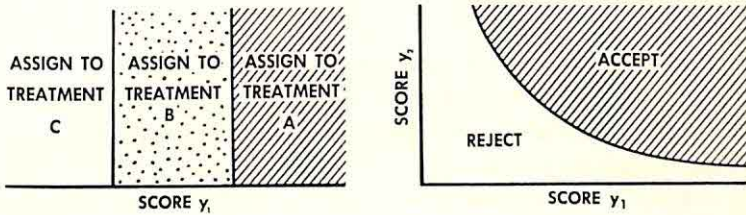


Figure 2. Placement and selection procedures

is where information is univariate. Even when information is obtained in terms of more than one score or dimension, it is a common practice to combine such multivariate information into a single composite score before making decisions. If scores on the same composite scale are used in making all decisions between treatments, the classification problem may be termed a placement problem (Figure 2). The most common examples are dividing students among sections to be taught at different rates, or using a trade test for coarse grouping of applicants. Use of a composite score to discriminate the brain-injured from the mentally deficient also exemplifies a placement decision. We shall see later that "measurement" problems can be considered as a particular variant of the use of tests for placement decisions.

Problems may also be differentiated according to whether or not rejection is allowed as one possible treatment, so that the person is eliminated from the institution. We can refer to these as selection problems; Figure 2 illustrates a selection strategy based on a non-linear combination of two scores. We shall give particular attention to the relatively simple problem of selection on the basis of univariate information, where all accepted men are treated alike.

Constraints upon decisions

Decision problems, whether concerned with classification, placement, or selection, may be distinguished on another basis: the presence or absence of certain constraints. Two types of constraint are common: (a) number of treatments per man, and (b) number of men per treatment. The great majority of institutional decisions call for one treatment per man. The man is sent to one training course or to one job, or given one diagnosis. In individual decisions

also, it is common to select one treatment at a time (for example, in choice of curriculum). Some decisions involve multiple treatments, however, especially where specific duties rather than job classifications are assigned. In organizing a ship, for example, a man may be assigned to a combination of duties: radar watches, radio watches and maintenance, and lookout duty. Likewise a student may need to select five courses to fill a program of studies. When an individual may be assigned to more than one treatment, the decisions with regard to the several treatments are usually not independent. The problem then becomes one of finding the best pattern of treatments.

In institutional decisions the number of men per treatment may be constrained by an absolute or relative quota. An absolute quota is set when there is a definite number of vacancies, whereas a relative quota calls for assigning to the treatment a certain proportion of the number tested. Instead of applying the constraints to each single sample, experience over a long period may be considered. Thus a minimal acceptance score may be determined which admits enough persons of good potentiality to maintain approximately the desired number of workers, students, etc. It is convenient to summarize a system of this character by translating the cutting score into a relative quota (one form of which is the familiar "selection ratio"). This quota states what fraction of a large population will be assigned to the treatment.

Not all institutional decisions are subject to a quota. Educational placement, in particular, may often be made without limiting the number of individuals who will receive any one treatment.

SEQUENTIAL AND SINGLE-STAGE STRATEGIES

We make a further important distinction between sequential and non-sequential strategies. The usual test theory assumes that a final decision is made on the basis of a battery of tests administered to all persons. This is a single-stage, i.e. non-sequential, decision procedure. It is possible, and probably more common in practice, to approach decisions sequentially. No irrevocable commitment of an individual to a treatment is made. Instead, after each stage of information gathering the individual is sent on to some further experience where new information about him is obtained. This new information may modify his ultimate disposition. The decision to reject a job applicant is typically final and not to be reconsidered, but a decision to accept the man is really a decision to investigate him further. His work on the job provides further data, and if he proves unsatisfactory he can be discharged.

Personnel decisions typically form a long chain. A student makes decisions about vocational choice in each successive year of schooling. Even after he has completed his training, he may shift into some other field or specialize

within his field. Choice points where he altered his plan may be readily recognized; there are also less conspicuous points where he considered a change of goal but did not make one. The institution is likewise making a sequence of decisions about the student's acceptability in the chosen career. Each time grades are given or he is admitted to a new level of training, the institution chooses between eliminating him and "investigating further". The college admissions office may view a decision to admit a borderline student as ending the decision process; but from the perspective of a dean considering the student's poor marks a year later, it was a decision to gather further information by observing his college performance.

After a test is given, one either can assign the person to a treatment or can decide to administer further tests. Broadly speaking, those further tests may include job tryouts, education and training programs, or even initiating psychotherapy and observing the patient's response. A distinction between "tests" and "treatments" is a convenience, but one of doubtful logical justification. Choice of any treatment which does not remove the individual from the hands of the decision maker is to some extent an "investigatory decision", since new facts can always be used to modify an earlier plan. A decision-making procedure which permits information at one point to determine what information will be gathered next is called a sequential strategy. It will prove important and profitable to extend test theory to encompass sequential methods.

A TAXONOMY OF DECISION PROBLEMS

The foregoing distinctions suggest a possible taxonomy for decision problems. Problems similarly classified are describable within a set of general principles. While a taxonomy adequately covering the entire range of testing problems would be hopelessly complicated, a small effort in this direction will serve to summarize the present chapter and to indicate how broad is the domain that concerns the tester. This will make particularly clear the fact that subsequent chapters barely begin the needed exploration of personnel decisions.

To classify any decision problem, we ask these questions:

1. (a) Are the benefits obtained from a decision evaluated in the same way for each person? or
(b) Are different values used in deciding about each person?
2. (a) Is the decision about each individual made independently? or
(b) Are decisions about various persons interrelated?
3. (a) Is each individual assigned to just one of the available treatments? or
(b) May he be assigned to multiple treatments?

4. (a) Is one of the allowable treatments "reject"? or
(b) Are all persons retained in the institution?
5. (a) Is the information used in univariate form? or
(b) Multivariate form?
6. (a) Are decisions final? or
(b) May one decide to obtain further information prior to final decisions?

These six questions define $2^6 = 64$ different patterns. While all of them might be described within one highly generalized model, such generality is likely to be confusing rather than clarifying. For this reason we shall discuss relatively specific problems within the 64 possibilities.

We can use a code pattern such as aaaaba to describe any decision problem in terms of the six questions. The symbol aaaaba would describe (1a) an institutional decision (2a) without a quota constraint, (3a) assigning a person to one treatment (4a) which may be "reject", (5b) on the basis of multivariate information (6a) obtained in a single stage of testing. The following chapters will be especially concerned with the patterns aaa... and aba..., that is, with institutional decisions where each individual is assigned to one treatment, and where a quota may or may not be employed. Within aaa... or aba... there are eight possibilities:

- ...aaa Reject decisions allowed; univariate information; single-stage testing. This is the common selection problem, considered in Chapter IV.
- ...aab As above, with sequential testing. This is the concern of Chapter V.
- ...baa No persons rejected; univariate information; single-stage. These placement and measurement problems are the concern of Chapter VI.
- ...bab Sequential placement, considered in Chapter VII.
- ...aba } Multivariate information; single-stage; rejection may or may
...bba } not be allowed. This is the classification problem considered in Chapter IX.
- ...abb } Sequential classification, considered briefly in Chapter IX.
...bbb }

Even this tabulation does not indicate all the distinctions with which test theory may be concerned. Further distinctions remain to be introduced in later chapters.

In distinguishing between so many different decision problems, this chapter makes clear the need for broadening existing test theory. Any test theory well developed at present deals with only a very limited area within our category system. The classical measurement theory of Kelley, Hull, and Gulliksen deals entirely with pattern aaabaa (multivariate information, where considered, being reduced to a linear composite). Cases involving "reject" decisions

present special characteristics. Some of these patterns (chiefly aaaaaa) have been treated in scattered papers, but only Brogden has done connected studies on them. We are forced to conclude that there is no "theory of mental tests" at the present time, although there are many fragmentary theories.

The present report shows some connections and differences between the various systems, and may therefore be regarded as a first step toward an integrated theory. There will probably always be a need for many alternative test theories covering different cases, however. While an abstract mathematical treatment can cover all or most of the range of personnel problems simultaneously, definite treatment of specific problems will usually be more comprehensible to the test user. The purpose of the overriding system is to make clear the differences among problems, so that misapplication of a sub-theory will not lead to incorrect practices.



3

CHARACTERISTICS OF DECISION PROBLEMS

All personnel decision processes can be characterized in the same manner. There is, in the first place, an individual about whom a decision is required, and two or more treatments to which he may be assigned. The decision is to be made on the basis of information about the individual.

The information is processed by means of some principle of interpretation, or strategy, which leads to either a terminal decision or an investigatory decision. A terminal decision ends the decision-making process by assigning the individual finally to a treatment. The outcome is his performance under that treatment. An investigatory decision calls for additional information, dictating what test or procedure will be used to gather that information. This then leads to a further decision. The cycle of investigatory decision, information gathering, and decision making continues until a terminal decision is made. The decision process is illustrated schematically in Figure 3. As was discussed in Chapter II, a decision is terminal only from the viewpoint of a particular decision maker. A treatment will often yield additional information about an individual which can serve to modify the treatment.

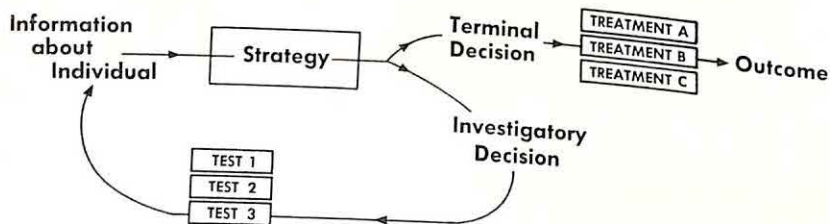


Figure 3. Schematic view of a decision process

STRATEGIES

A strategy (or decision function) is a rule for arriving at decisions. A strategy must state what the decision maker will do in any possible contingency (cf. 36). Thus a rule for college admission may say, "Any graduate of an accredited high school will be admitted; other applicants will be expected to pass certain tests of General Educational Development". This is equivalent to the following formal strategy:

Given information that individual i is a graduate, make the terminal decision "accept" with probability 1.

Given information that i is not a graduate, make the investigatory decision "Give GED tests" with probability 1.

At the second stage, given the further information that i 's score on GED is Pass, make the terminal decision "accept" with probability 1. Given information that i 's score on GED is not Pass, make the terminal decision "reject" with probability 1.

A strategy consists of a set of conditional probabilities. Probabilities employed in a strategy need not be restricted to 1 or zero. Given certain information about the individual, the probability of each decision is specified. The probabilities that constitute a strategy can be written as a matrix. Strategy matrix 1 describes each stage of the admission rule stated above, and matrix 2 combines the two stages. Every entry is of the form $p_{d/y}$ -- the probability of making decision d , given information y . Matrix 2 makes clear that a strategy specifies the decision to be made in each possible contingency. Under the present strategy such a condition as "graduate, non-pass GED" cannot arise; if the test were routinely given to every applicant, however, the decision rule would need to provide for this contingency also.

The word strategy suggests that a conscious policy guides decisions. Even if choices are based on habit or some chance mechanism rather than policy, they can be described in a matrix. Each entry states what proportion of the

Information Category (y)	Decision		
	Accept	Reject	Continue Testing
First Stage:			
Graduate	1	0	0
Non-Graduate	0	0	1
Second Stage:			
Non-Graduate, Pass GED	1	0	0
Non-Graduate, Non-pass GED	0	1	0

Strategy Matrix 1. Plan for two stages of decision regarding college admission

Information Category (y)	Decision				
	Accept	Reject	Give GED, Accept	Give GED, Reject	Give GED, Continue Testing
Graduate	1	0	0	0	0
Non-graduate, Pass GED	0	0	1	0	0
Non-graduate, Non-pass GED	0	0	0	1	0

Strategy Matrix 2. Same plan, combined form

time the decision maker makes each choice, under each set of possible conditions (information). The matrix describing a decision maker's practice must be based on tabulation of a large number of decisions. The probabilities recorded will often be other than 1 or 0. For example, suppose we examine recommendations of an interviewer for a private college where all applicants take a certain scholastic aptitude test. A tabulation of his decisions might report his decision pattern or "strategy" as this matrix:

	Decision	
	Accept	Reject
High-school graduate, SAT above 70	.80	.20
Not high-school graduate, SAT above 70	.80	.20
High-school graduate, SAT below 70	.20	.80
Not high-school graduate, SAT below 70	.10	.90

Every entry is a conditional probability. This interviewer evidently gives much more weight to SAT than to the fact of graduation. The table tells us that he accepts some graduates below 70; we do not know whether this is due to information not used in classifying entries for the table, or to sheer inconsistency in decision making.

A description such as this often reveals contingencies of which the decision maker himself is unaware. His strategy, that is, may be implicit rather than explicit, and may actually depart from some stated policy. Description of actual strategies is an important way to study industrial interviewing and rating, clinical diagnosis, and social perception. It is known, for example, that diagnostic teams in different hospitals report strikingly different proportions of certain pathologies among their intakes. This is apparently to be accounted for more by biases in the decision process than by differences in the intake populations. The overall proportions summarize some aspects of the "strategy" of the diagnostic team. More refined research would relate the probabilities to the categories of diagnostic information.

In any situation, some strategies are better than others. Just what constitutes a "best strategy" depends upon some subtle questions of evaluation, but

once a way of evaluating outcomes is agreed on, the benefit from various strategies can be compared. If the decision maker aims to maximize expected utility, it is always better to assign persons in the same information category to the same treatment than to distribute them without further information. Sometimes a quota forces one to divide persons with the same observed characteristics, but otherwise in an ideal strategy matrix for institutional decisions the entries will always be 1 or 0.

Tabulation of actual strategy matrices reveals ways in which the decision-making practice departs from the ideal. Within a given category, the decision maker often differentiates, for example recommending different treatments for persons whose score patterns are negligibly different. He may believe that this differentiation is based on additional information not represented in the scores, for example, impressions gained from interviewing. Unless such added information has considerable validity, the judge who differentiates within an objectively defined group may actually impair the utility of his decisions by introducing random variation (26; 53, p. 117). Once systematic errors in strategy are identified, they can be corrected by retraining.

EVALUATION OF THE DECISION-MAKING PROCEDURE

In evaluating a strategy for making decisions, one asks such questions as these:

1. Does this procedure arrive at the best decisions possible, with this body of information?
2. Would gathering some other (or additional) information permit better decisions?
3. How much difference is there in the goodness of decisions arrived at by any two procedures?

Evaluation of a strategy involves evaluation of possible outcomes and the prediction of possible outcomes; these require separate consideration. First, it is necessary to state what value the decision maker places on each possible outcome. Then if the outcomes from a given strategy can be predicted, it is possible to evaluate the utility of the strategy.

Outcomes

The outcome consists of all the consequences of a given decision which concern the person making the decision (or the institution he represents). What outcome will result when the chosen treatment is applied depends on the characteristics of the individual and ordinarily on further unspecified situational variables. For example, the outcome of a decision to accept a student depends on unmeasured motivational factors on the one hand, and on the particular courses and instructors he selects on the other.

The decision problem would be enormously simplified if the decision maker could anticipate the actual outcome for each person under any treatment. A systematic comparison would establish the best assignment. Having only fallible information, the decision maker can at best predict the probability distribution of outcomes, or state the expected outcome over many similar decisions.

Prediction of outcomes

An outcome is described in terms of some criterion or set of criteria. In industry, outcomes of hiring include the man's hourly production, his spoilage, his length of stay with the company, his effect on the morale and tenure of other employees, etc.

Empirical results from previous cases are required to determine how the information (y) is related to the criterion (c). This experience provides a validity matrix. There is one validity matrix for each treatment. In general form, it consists of a set of conditional probabilities $p_{c/yt}$. The criterion

Information Categories (y)	Criterion States (c)		
	1	2	3
1	$p_{1/1t}$	$p_{2/1t}$	$p_{3/1t}$
2	$p_{1/2t}$	$p_{2/2t}$	$p_{3/2t}$
3	$p_{1/3t}$	$p_{2/3t}$	$p_{3/3t}$

Validity matrix for treatment t

states may be simple judgments ("successful", "unsuccessful"), numerical scores, or complex categorizations of behavior. The validity matrix is like a contingency table, save that the latter is customarily stated in terms of joint probabilities or frequencies. In our computations, we shall assume that if the operational criterion is multidimensional it can be reduced to a single dimension by the application of appropriate weights. The validity matrix then takes the form of a multivariate distribution of test and criterion scores.

Valuation of outcomes

Each information category leads to a different expected distribution of outcomes for a particular treatment. To compare alternative strategies for assigning individuals in a particular information category to a treatment, one must judge the desirability of each possible outcome. To evaluate outcomes cardinally requires assigning a value e_c to each possible criterion score. We shall assume here that such evaluations are expressed on an interval scale of utilities, but they may vary with c in any manner. To take a specific example, consider a criterion scale reporting the proportion of perfect objects produced

by each machine operator. The higher this quality measure, the more valuable the operator is -- but the benefit may not be a linear function. At the extreme upper end, a slight rise in quality of output may permit elimination of routine inspection. If so, a 2% rise in quality from 94% to 96% would represent a great gain. Where the quality is low, on the other hand, complete inspection is required, and the operator whose score is 76% yields negligibly more benefit than one at 74%. Practical difficulties of assigning values to outcomes will be discussed in Chapter X.

The evaluated outcome may be referred to as a payoff. We can determine the expected payoff from an individual in a particular category y_i by weighting the value (e_c) of each outcome by its probability ($p_{c/y,t}$), and summing over all outcomes. The cost of a test or any other procedure for gathering information must be expressed in utility units and deducted from the expected payoff.

Expected utility from a strategy

As we have formulated the problem, a strategy is only to be evaluated by its total contribution when applied to a large number of decisions. "Expected payoff from an individual" has little meaning unless one can average actual payoffs over a large number of individuals in the same information category. For this reason our model as developed here does not yield results interpretable with respect to individual decisions.

If we know the distribution of scores y_i in the population tested, the expected net utility for a large number of decisions is determined simply by adding the expected payoff for each y_i , weighted by the probability of that score. This can be stated algebraically, using the following symbols:

- U = utility of the set of decisions
- N = the number of persons about whom decisions are made
- y = information category
- t = treatment
- c = outcome
- e_c = value of outcome
- C_y = cost of gathering information

$$U = N \sum_y p_y \sum_t p_{t/y} \sum_c p_{c/yt} e_c - N \sum_y p_y C_y \quad [1]$$

The p_y describe the assumed y distribution, the $p_{t/y}$ are the entries of the strategy matrix, and the $p_{c/yt}$ are the entries of the validity matrix. C_y may or may not vary from score to score. Whatever strategy gives the greatest value of U is to be preferred.

PAYOFF FUNCTIONS RELATING UTILITY TO APTITUDE AND TREATMENT

If we are dealing with a single treatment, and prediction is based on continuous test scores, we may conceive of a function relating expected payoff e to score y . This payoff function states how much benefit a person with any given score is expected to contribute through production or learning or other outcome.

In selection there is sometimes only one possible treatment for accepted persons. More frequently, however, a variety of treatments might achieve the same ends. Thus, different curricula might be applied to attain certain educational objectives. Within an industrial organization, the responsibilities of various jobs might be modified.

Each treatment will have its own payoff function relating expected payoff to score. The payoff function, while presumed to be continuous, may take a variety of forms. Figure 4 shows possible payoff functions under different

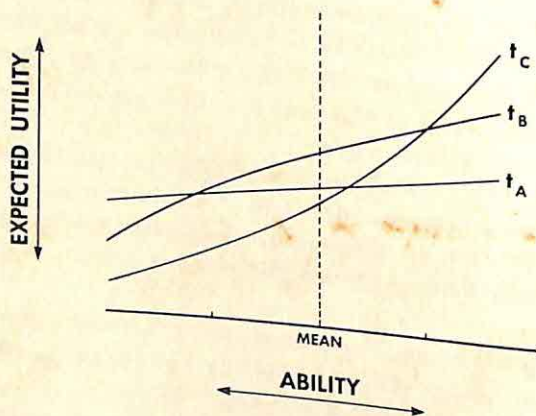


Figure 4. Three payoff functions

treatments, with univariate test information. According to the figure, very able men will perform best on treatment C, and least well on treatment A. With very low scores the situation is reversed, and A is the optimum treatment.

An illustrative situation is to be found in technical training. The curriculum could proceed at a rapid pace, assuming considerable knowledge of mathematics, technical vocabulary, and basic concepts on the part of the student. With a superior student such a method would produce rapid gains, but a weak student would progress slowly. More leisurely instruction which fills gaps in fundamentals and demands less mathematical understanding would permit the weak student to make maximum gains, but it would not develop the well-prepared student rapidly.

Little empirical information regarding payoff functions is presently available. For any single treatment, difficulties of prediction rarely permit us to fit any function more complex than a linear regression formula. Conventional comparisons of treatments show mean differences between payoff under two production methods or two instructional techniques. These results indicate only that one treatment yields higher outcomes for persons who are on the average like the experimental subjects, and do not yield payoff functions. Such an experiment would indicate, regarding Figure 4, that treatment B is best. Significant interactions between ability and treatment are reported in some studies, but the data are generally inadequate to plot payoff functions. The scattered experimental evidence (see 2, for example) warrants the belief that payoff functions for different treatments will differ in slope as well as in mean value.

Even more to the point is the fact that use of tests for educational placement implicitly assumes the existence of such different payoff functions. Our model therefore is merely a formal statement of what is everyday accepted without question, and examining its implications for testing is undoubtedly important. The realization that this assumption underlies practice is in itself an advance, because it makes clear the need for research on payoff functions. But let us first defend the claim that this assumption is currently made.

Dividing a group on the basis of some univariate information and assigning each segment to a different treatment must be justified by the belief that greater payoff will be obtained in this manner. With two treatments, there are three possibilities.

1. The payoff functions are identical, in which case placement is valueless since the treatment to which a man is assigned makes no difference in payoff.
2. The payoff function for one treatment is uniformly higher than the other for all score levels. In this event it is unwise to divide the group. All persons should be given the superior treatment unless institutional constraints require both treatments to be used (for example, where there are too few therapists to give psychotherapy to all patients).
3. The payoff functions intersect somewhere within the score range. If this is the case, as in Figure 4, it is profitable to assign men to different treatments, so that each receives the treatment which "gets the most out of him".

Dividing students into sections to receive instruction at different rates clearly assumes the existence of condition 3, for there is no administrative necessity to vary the instructional pace from section to section.

The assumption that payoff functions do intersect, making placement profitable, is consistent with available theories about instruction. A person who

lacks readiness to profit from one experience may be able to learn from another. Theories about response to therapy and about job performance likewise support an expectation that payoff functions will intersect. We therefore anticipate that empirically determined payoff functions will often be like those of Figure 4. Accepting the concept of intersecting payoff functions will allow us (in Chapter VI) to raise fundamental questions regarding the construction and validation of placement tests. We shall also use the concept as a base for studying certain selection problems.

Multivariate response surfaces

The problem of the tester may be further clarified if we think of the payoff function as a response surface (9) in multidimensional space. Under any one treatment, the outcome for a given person might be presumed to be exactly predictable, if enough is known about his characteristics and the treatment is precisely controlled. Actually, some variance will always be unpredicted, and therefore one can only describe the expected payoff for a person having certain characteristics. It is convenient to think of the measured characteristics as expressed in terms of orthogonal factors $s_j = s_1, s_2, \dots$. Any test score may be resolved by factor analysis into a linear composite of certain s dimensions. The person's characteristics are described by the pattern $\tilde{s}_i$. For each pattern $\tilde{s}_i$ there is an expected payoff e_{it} . The payoff function is a response surface continuous in the s_j . A different surface is expected for each treatment.

In any decision problem, we are actually concerned chiefly with a subset of the various s 's, those which are related to payoff. For any set of treatments under consideration, the test scores may be so factored as to extract those dimensions which account for variance in payoff under at least one treatment. Payoff under any treatment depends on one or more of these aptitude dimensions. Any other of the test-score dimensions will have zero coefficients in payoff functions for all the treatments. Our argument to this point has emphasized that one may choose that one of several distinct treatments for which e_{it} is greatest, as estimated for an individual from whatever test information is available.

Quite a different response surface may be considered if we regard treatments as varying continuously. Treatments may vary along a large number of dimensions, which might be designated $w_1, w_2, \dots$. The classic example is found in chemistry, where temperature, pressure, and concentration of some reagent may be varied. The yield of the reaction will vary accordingly, and a continuous response surface may be fitted to observed points. Experimental psychologists likewise are accustomed to think of factors in a treatment as continuous variables; time between trials, proportion of reinforcements, amount of ego-involvement, and discriminability of cues, for example, affect various forms of learning in a presumably continuous manner. In personnel psychology, precise control of treatments is difficult and very often

the decision maker is confined to a few alternatives. But in principle the treatments given a student may vary in a continuous manner. Such variables in an instructional situation may include the amount of supportiveness of the instructor, the pace of instruction, or the extent to which the instruction demands mathematical ability. In psychotherapy, the treatment may vary in warmth, depth of interpretation, and extent of medical supplementation. If the concept that treatments vary continuously is accepted, for individuals having the characteristics $\tilde{s}$ there is a response surface $e_{st} = f(\tilde{w}_t)$. This surface presumably has an absolute maximum (and perhaps local maxima).

If the relation of payoff to the treatment parameters were independent of the person's characteristics, then such parameters would be of little interest to the personnel tester. The same treatment would be best for all individuals. There is often an interaction, however, among individual characteristics and treatments, so that the location of the maximum payoff differs for different patterns $\tilde{s}$.

The personnel worker is usually concerned with choosing individuals to fit a fixed treatment. In clinical and guidance testing, however, it is the individual that has been fixed and the treatment is to be selected from many possibilities. The task of the clinician appears to be to adjust the treatment as nearly as possible to the optimum for the individual. He can ordinarily use his information to select from among an almost infinite variety of treatment conditions that one pattern which seems best for the person. It is true that present sources of information are a poor basis for making such adaptations, but this indicates only that gains from adaptation are likely to be limited at present. Certainly the physician does not hesitate to compound drugs, rest, exercise, diet, and bedside encouragement in an individualized formula. The same flexibility is ultimately open to the teacher and the psychotherapist, but their supporting science is less adequate at this time.

The response surface linking payoff to treatment factors has been ignored in test theory because it plays little part in traditional personnel decisions. Assigning men to fixed categories, or predicting their scores under a single treatment, is all that the industrial and military psychologist has attempted. But even in personnel classification adaptation may be possible. One may vary such important conditions of a job as amount of on-the-job instruction, amount of supervision, and pacing of work. Introducing radical changes in degree of responsibility or amount of automatic control makes changes in payoff even more likely. So long as one can expect to employ men of a given quality, one should set the treatment so as to maximize their payoff. Within the limits of practicality, a change in quality of men calls for adaptation of treatment. Thus in all personnel testing there may be need to consider payoff as a function of treatment variables.

Considering both types of variables affecting payoff, we see that $e_{it} = f(\tilde{s}_i, \tilde{w}_t)$. Since both $\underline{s}$ and $\underline{w}$ are (in principle) multivariate, the problem cannot be easily represented geometrically. Algebraic analysis is complex but not

impossible. Rather than attempt a comprehensive study of such payoff functions, we shall mainly examine problems involving a single aptitude dimension s . We shall restrict ourselves also to a subset of treatments whose payoff functions with respect to s vary in a continuous manner. These radical restrictions permit exploration of only a corner of the total problem, but a great deal is to be learned even from such a limited case.

The foregoing considerations lead us to make a distinction between decision problems involving "fixed-treatment" and "adaptive-treatment" conditions. In the former, the treatment is specified a priori, without regard to information from the particular persons tested. For example, a job or curriculum is defined and an attempt is made to select the individuals who will perform best in it. In adaptive treatment, however, the choice of men and choice of treatment are both made in the light of test data. Thus the personnel manager may be allowed to reduce the difficulty of a job when he has to fill vacancies with a poorer quality of applicant. Speaking generally, the aim of adaptive treatment is to choose the treatment for an accepted person in order to fit his aptitudes as well as possible, or to fit the group of accepted persons as a whole if all must be given the same treatment. So long as the payoff function for any admissible treatment is linear and positive in slope, the same men will be selected under either fixed or adaptive conditions. As Chapters IV and VI will show, however, the contribution of testing is different in the two conditions. Chapters VII and VIII will demonstrate that the testing procedure itself may profitably be altered as one shifts from fixed to adaptive-treatment assumptions.

Assumption of first-degree payoff function

In the mathematical developments of the following chapters, we shall assume first-degree functions relating expected payoff to aptitude under any treatment. When a single aptitude is involved (as when a single test score is used), there is a linear payoff function. This simple relation has been used by many previous investigators (e.g., 10, 20, 55).

If one wishes instead to consider payoff functions of higher degree, a very large number of possibilities would be available, and present empirical knowledge gives no basis for choosing any one of these as especially realistic. Fortunately, a linear relation serves as a good approximation to many higher degree functions, especially when there is considerable unpredicted criterion variance.

4

SELECTION DECISIONS WITH SINGLE-STAGE TESTING

Chapter IV examines the most familiar of personnel decisions, selection on the basis of a single test or composite score. The most common examples are industrial selection, admission of students, and psychiatric screening of recruits. Typically, a constraint is placed on the number or the minimum quality to be accepted.

A discussion of selection decisions at this point has many advantages, even though not all the results to be presented are new. Stating a familiar problem in terms of our model will clarify utility analysis, and lay the groundwork for the consideration of other personnel decisions.

THE INTERPRETATION OF VALIDITY COEFFICIENTS: EARLIER WORK

Evaluating the benefit obtained from tests is of considerable practical interest. Many of the studies have been stimulated by the "public relations" problem of convincing business management or military authorities that the benefits from testing programs justify their cost. Professional workers, in planning testing programs, must decide which tests to use, and how many; and this requires balancing costs against estimated benefits.

The contribution of a testing program depends on the importance of the decisions to be made, the selection ratio, and possibly other characteristics of the situation. The situation being fixed, the cost and validity of the test determine whether it should be used and how much is gained by using it. Thus assessing the value of the testing program requires a study of the benefits associated with any level of test validity.

The relation of benefit to validity has long been regarded as an important question. Interpretation of validity coefficients is an important topic in the training of test users, as is evidenced by its ubiquitous appearance in professional texts on testing and also in literature on testing aimed at a more general audience. As at least four different formulas for interpreting the validity

coefficient have been given, we face the problem of choosing between them or of reconciling them.

The validity coefficient is usually defined as the correlation of test score with outcome or criterion score. We shall use r_{ye} in our discussion because the decision maker is interested in predicting the evaluated outcome, that is, the outcome expressed in utility units.

The rule having the longest history translates the validity coefficient r_{ye} into the "index of forecasting efficiency", $E = 1 - \sqrt{1 - r_{ye}^2}$. The prominence of this index and its close relatives is attributable to the emphasis placed upon them by Truman L. Kelley (46) in his Statistical Method (1923) and by Clark L. Hull (43) in Aptitude Testing (1928). The index compares the standard error of scores predicted by means of the test to the standard error of chance estimates. The proportionate reduction of the standard error is taken as a measure of the value of the test.

The "coefficient of determination" r_{ye}^2 provides a second scale for evaluating a test. This coefficient expresses the ratio of predicted variance in payoff to the total variance. Although it appeared in Hull's book (43, p. 267), it has become prominent only much more recently.

According to the customary interpretations based on these formulations, a high validity coefficient is necessary to obtain substantial benefit from testing. The index of forecasting efficiency describes a test correlating .50 with the criterion as predicting only 13% better than chance; the coefficient of determination describes the same test as accounting for 25% of the variance in outcome.

Such generalizations do not recognize that the value of a test varies with the particular decision for which it is to be used. This point was first emphasized by Taylor and Russell (64) in 1939, following an earlier statement by Thurstone (66, p. 59). They examined what proportion of employees are likely to be "satisfactory" before and after selection by means of a test. This proportion depends on the percentage of applicants accepted (selection ratio). A test with validity .30 permits substantial improvement when the selection ratio is low, and under some conditions a test with validity .80 gives virtually perfect selection. The Taylor-Russell report challenged conservative traditions in the measurement field: "The wide-spread acceptance of such measures [as E and r^2] as the correct way of evaluating correlation coefficients has brought about a considerable pessimism with regard to the validity coefficients which are ordinarily obtainable when tests are tried out in the employment office of a business or industry or in an educational institution. . . . Very considerable improvement in selection efficiency . . . may be obtained with small coefficients."

Brogden (10) evaluated decisions directly on a utility scale to interpret the validity coefficient, and concluded that in selection the gain from use of a

test is linearly related to the validity, regardless of the selection ratio. He argues that a test of validity .50 gives 50% of the improvement that would result from using a perfect test. This result is contrasted with an interpretation based on the coefficient of forecasting efficiency which Brogden quotes from an outstanding text on measurement published in 1936: "Tests with a coefficient of validity less than .5 are practically useless except in distinguishing between extreme cases." The linear relation has also been discussed by Cochran (20), and in modified form by Richardson (55), Jarrett (45), and Brown and Ghiselli (16).

The Brogden generalization, like that of Taylor and Russell, encourages far more extensive use of tests than do the interpretations in terms of r^2 or $\bar{E}$. Wesman (70) rationalizes this seeming discrepancy by pointing out that an error of prediction regarding a superior man who is accepted does no harm. So long as it is wise to hire both A and B, it does not matter that the test overestimates A's production, and underestimates B's. Only errors which "cross the borderline" so that a man is hired who should not be, or vice versa, impair the utility of the decisions. The analyses disagree because the coefficients $\bar{E}$ and r^2 assign evaluations differently than do the more recent analyses. In computing $\bar{E}$, the loss to the institution is regarded as proportional to the discrepancy between predicted and actual performance; and in the coefficient of determination, as proportional to the square of this discrepancy. But a man's contribution in a particular setting depends on his performance, not on the prediction made about him. Errors in prediction are costly only when they cause errors in decision. This is recognized by the Taylor-Russell treatment and, in slightly different ways, by its successors.

The value of a test can be stated only in terms of the specific type of decision problem, the strategy employed, the evaluation attached to the outcome, and the cost of testing. We shall use the utility model to derive this relationship for simple selection decisions, and later for a variety of other conditions. It will be increasingly clear that the linear interpretation of the validity coefficient, or some minor departure from it, is appropriate for simple selection, but that other indices must be chosen in other decision problems. Brogden's insightful statement indicates why interpretation of validity coefficients presents so much difficulty:

In general, it is probably true that statistical formulas are not developed with the primary objective of providing interpretations most meaningful for a research worker having problems peculiar to a given area of research. The formula is more apt to be developed as an expression of certain mathematical relationships. In the derivation, assumptions -- often highly limiting in nature -- are introduced as necessary to the development of a given formula. Applications are sought at a later date. Very often it is found that the assumptions are so restrictive that the coefficient can legitimately be used in only a small proportion of certain types of application. In other instances the coefficient may have legitimate application but may not provide the interpretation needed. (12, p. 170)

BASIS FOR JUDGING GAIN FROM TESTING

It is customary to interpret the value of tests in terms of improvement over chance. "Percentage of possible improvement" is frequently reported, the maximum possible improvement being the gain that would result if a test of validity 1.00 were used. This is a relative measure of gain. Occasionally writers have examined absolute gain, using some utility unit such as the Brogden-Taylor "dollar criterion" (14). Richardson (55) introduced still a third variant, obtaining an efficiency ratio by dividing the absolute utility after testing by the absolute utility from unselected men.

Chapter II suggested that expected outcome under any strategy should be evaluated on an absolute utility scale. The value of a strategy involving a test can be compared to the utility from the best strategy not using the test ("best a priori strategy"). The difference is the gain from testing, and is positive or negative depending on the cost of the test. We shall not translate this difference into a ratio.

The proposal to consider the "best a priori strategy" as a baseline may appear to be only another way of referring to "improvement over chance". But this rephrasing will perhaps significantly alter judgments about the value of a particular test. Most important, it leads to the conclusion that the test-criterion correlation based on a population of completely unselected applicants is often not the proper validity coefficient for judging the value of testing.

A priori population

The usual validity coefficient reported for a test is the zero-order correlation relating test to criterion in an unselected group of applicants. It has perhaps been insufficiently recognized that the Taylor-Russell paper and others subsequent to it do not necessarily interpret this validity coefficient. The Taylor-Russell discussion presupposes a validity coefficient based on present employees. This population is likely to have been screened using information other than the test. It is to this population that the selection ratio will be applied. Jarrett requires that the validity calculation be based on individuals who "would have been put to work", i.e., who would have met any requirements already in use. This definition is consistent with the Taylor-Russell argument, and we also adopt it.

We use "validity" subsequently to refer to a correlation computed on men who have been screened on whatever a priori information is in use and will continue to be available. This pre-screened group is referred to as the a priori population. Such a validity coefficient is lower than that based on all applicants, the reduction being due to truncation of range through selection. Any zero-order coefficient based on unselected persons must therefore be reduced, before applying the interpretations derived below.

Best a priori strategy

Even when the a priori population is defined as suggested by Jarrett, chance selection is not necessarily the alternative strategy with which testing is to be compared. The decision maker who does not test may use interviews or other available information to provide some basis for a decision. Thus, military recruits could be screened with better-than-chance efficiency by utilizing recorded information about civilian job experience and grade of school completed, instead of tests. Tests will improve decisions, but not as much as is implied by the phrase "improvement over chance". Similarly, if testing were impossible in educational placement or selection, one would not fall back upon a chance decision, because previous school records could be used with substantial predictive validity.

Tests should be judged on the basis of their contribution over and above the best strategy available, making use of prior information. The manner in which this comparison should be made depends on our contemplated use of that information. At least three possibilities should be considered.

1. All prior information is used and will continue to be used for pre-screening of applicants. In this case, validity coefficients should be calculated on the pre-screened group. The a priori strategy would be chance selection from this group to fill the quota.
2. Decisions will be based either on the test or on certain other information such as school records. One or the other will be used, not both. In this case, the difference in utility between the two strategies, taking cost of each into account, would indicate the advantage of testing.
3. Previously available information and testing information will be combined into a single score or pattern of scores on which decisions will be based. In this case, the gain due to testing is the difference between the utility obtained by the composite scoring technique and the utility obtained using only the previously available information.

Contribution of various types of tests

Evaluating a test in the light of its independent contribution to utility might reshape many of our present testing policies. Conrad makes this demand regarding validity information in test manuals:

... we ought to know what is the contribution of this test over and beyond what is available from other, easier sources. For example, it is very easy to find out the person's chronological age; will our measure of aptitude tell us something that chronological age does not already tell us? It is also easy, in some cases, in a local community, to obtain a person's school record. If that is so, then the question is, Does the intelligence test or the measure of aptitude tell us anything that is not already told by the previous information? The independent contribution is certainly something which should definitely be known, but very seldom is it revealed to us in the information which test publishers provide. (21, p. 65)

High validity coefficients have most often been attained for tests which are essentially work samples of their criteria. A typing test given an applicant, a test of mathematical ability given a prospective engineering student -- these tests offer good validity because they reproduce the criterion situation in miniature. Where such a work sample might be used, a record of past performance is also a valid basis for decisions, and is often routinely available. Because school marks are a good predictor of subsequent school marks, a scholastic aptitude test contributes much less to decisions than its validity coefficient suggests.

Per contra, tests having rather low zero-order validities often deal with attributes that are hard to test, and also hard to assess from any non-test data. Tests of leadership, social and emotional adjustment, or creativeness rarely boast validity coefficients beyond .30. But they predict qualities where useful alternative evidence is almost never available. Their contribution to utility in these situations might therefore be higher than the contribution of, say, the group intelligence test in decisions where it is relevant, after due subtraction in both cases for overlap with a priori information.

Many other examples may be given. In the classroom, growth in some types of skill and knowledge can be assessed by the teacher in her daily contacts with pupils through assignments, class discussion, and observations. These data, if cumulated with reasonable care over several months, will have sufficient validity that a test can improve the teacher's decisions by only a small increment. Other objectives of considerable importance are quite difficult for the teacher to judge, partly because opportunities for observation are scattered and partly because the criteria for judgment are difficult to formulate. These objectives include reasoning habits, attitudes, creativity, and application of knowledge. A test in one of these areas, even though it has severely limited validity, may yield more insight than a traditional achievement test which largely confirms what the teacher already knows. Generally, the more difficult it is to prepare a valid test for an objective, the more difficult it is for the teacher to make judgments about it without the test. Here also, then, the contribution of the test over and above available bases for the same judgment should be the criterion of its worth, rather than some measure of its accuracy, standing alone.

FIXED-TREATMENT SELECTION

In selection one decides between accepting an individual into an institution and rejecting him. The institution may be a university, an industrial plant, a military service, etc. In general, the aim of the decision maker is to accept those individuals whose expected payoff is highest, within the limits placed upon him by the available information and by whatever quota he must fill. In fixed-treatment selection, individuals are chosen for one specified "treatment" which cannot be modified.

For men accepted, a certain benefit accrues to the institution. We shall assume that this payoff has a linear regression on test score. A decision to reject a man will mean in most instances that he has no further contact with the institution. We may therefore regard the outcome of such a decision as having a value of zero. We shall also assume that test scores are normally distributed, with zero mean and unit standard deviation. This implies that the expected payoffs for randomly selected men are normally distributed.

This assumption of normal distribution of benefits may not be realistic. Standards set by management often place a lower limit on production which would cause the distribution to be skewed negatively. Voluntary limitation of output in a work group skews the curve in the opposite direction. And thorough standardization of a task eliminates variability entirely among those who remain on the job. Specifying a particular distribution permits a more complete exploration of functional relations, but generalization to other distributions is hazardous. Formulas derived under the normal assumption would need to be modified for any specified non-normal distribution of payoffs. Our general conclusions regarding the relationship of validity to utility depend only on the nature of the hypothesized payoff function. Many conclusions regarding particular optimal strategies, however, depend in addition upon the normal assumption. Brogden (10, 11) has examined thoroughly the effect on utility of departures from normality.

Linear relation of utility to validity

The net gain in utility per man tested from selection for a fixed treatment is linearly related to the validity of the test, under the assumptions stated above. As Appendix 1 demonstrates,

$$\Delta U = \sigma_e r_{ye} \xi(y') - C_y \quad [2] \quad (1.8)^*$$

C_y is the average cost of testing one person, r_{ye} is the correlation of the test with the evaluated criterion in the a priori population, and σ_e is the standard deviation of this payoff. y' is the cutting score on the test, and $\xi(y')$ is the ordinate of the normal curve at that point. In this expression, $\sigma_e r_{ye}$ is the slope of the payoff function relating expected payoff to score. It would be possible to employ here, as we do later, the concept of an aptitude s intervening between score and payoff, such that $r_{ye} = r_{ys} r_{se}$. This would make the development more closely comparable to that for adaptive treatment, but the argument would be less straightforward.

It is particularly important to note that the slope of the payoff function is influenced by several factors, since conventional methods place almost exclusive

*Where double numbers are provided, the former is the identifying number of the equation in the text, and the second locates it in the Appendix. (1.8) indicates the eighth equation in Appendix 1.

emphasis on the test-criterion correlation. An increase in correlation leads to an increase in slope, but slope also depends on the spread of criterion scores (which may vary for different treatments) and on the value associated with one unit on the criterion scale. The relation of gain in utility to selection ratio and validity is shown in Figure 5. On the vertical axis of this and subsequent

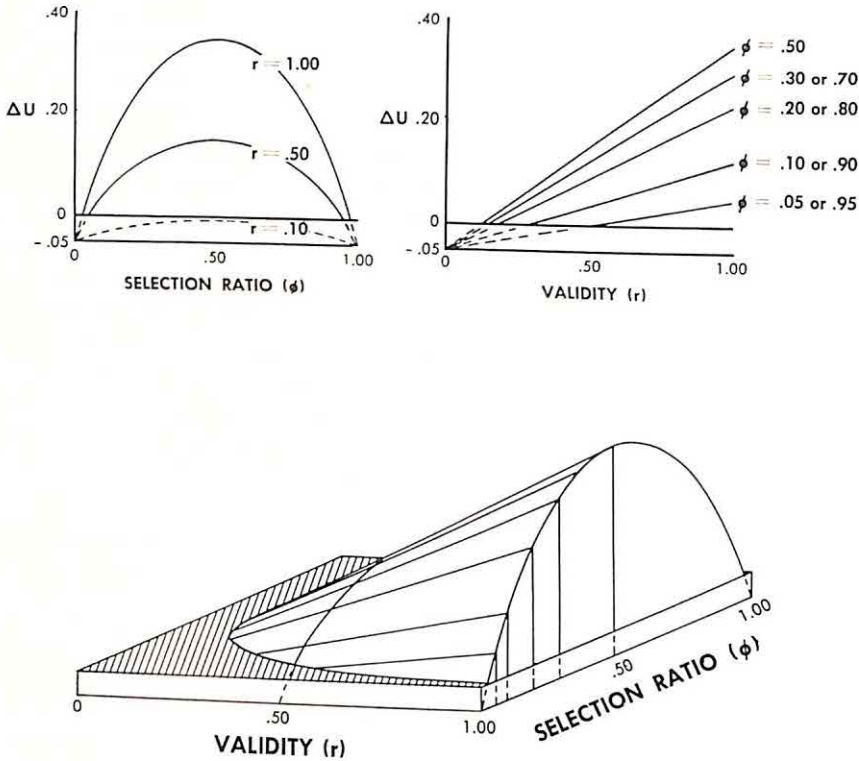


Figure 5. Relation of gain in utility to validity (r_{ye}) and selection ratio in fixed-treatment selection

figures $\Delta U/\sigma_e$ is plotted rather than ΔU . This does not distort the relations depicted. For this illustration, C_y is set equal to $.05\sigma_e$. This cost is probably unrealistically large, but such a cost aids in depicting relationships clearly. The shape of the surface would change only slightly with smaller values of C_y . In no case can $\Delta U/\sigma_e$ become larger than .399 as r increases and cost decreases.

Most of the conclusions implicit in [2] were originally pointed out by Brogden (10, 11, 12). Utility is a linear function of validity, and if cost is zero, is proportional to validity. With fixed-treatment selection and a linear payoff function, we may reject the interpretation of the validity coefficient in terms of r^2 or $1 - \sqrt{1 - r^2}$. Contrary to the Taylor-Russell results, the linear relation holds at all selection ratios.

Effect of σ_e on utility. A utility model brings in the important parameter σ_e . For any one treatment σ_e is constant, and indicates both the magnitude and practical significance of individual differences in payoff. A decision maker selecting men for many different assignments is frequently able to test for only a few of them, and must decide where tests can make the greatest contribution. The importance of an assignment can justify using a test of low validity (10, p. 71). Gain in utility depends on the product $\sigma_e r_{ye}$. Therefore a test of validity .30 for one decision may be more beneficial than a test having validity .60 for some other selection decision, if σ_e for the former decision is at least twice as large. A large σ_e is an indication that individual differences on the criterion in question have large practical importance. Tests for important decisions which fall far short of the ideal predictor may be much more worth using (and improving) than tests which give excellent guidance in making minor decisions.

Effect of cost on utility. Cost of testing should be taken into account along with validity and σ_e in deciding which test to use for a particular decision. The utility contributed by test 1 is greater than that from test 2 whenever

$$\sigma_e r_{y_1 e} \xi(y') - C_1 > \sigma_e r_{y_2 e} \xi(y') - C_2$$

Test 1 is preferable whenever

$$r_{y_1 e} - r_{y_2 e} > (C_1 - C_2) / \sigma_e \xi(y')$$

The difference in costs between two competing tests will not often be large

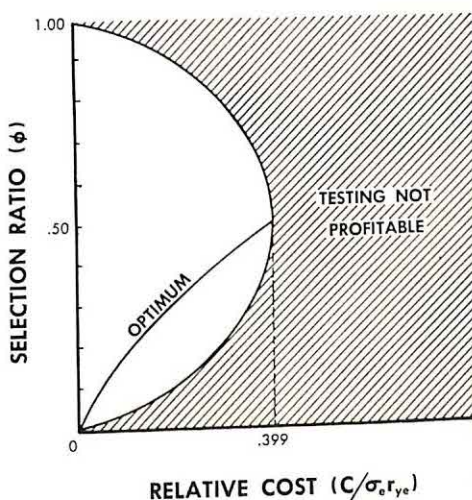


Figure 6. Utility as a function of selection ratio and cost

enough to justify use of the less valid test, but at some point (see Chapter VII) increases in test length can increase cost enough to offset accompanying gains in validity.

Effect of selection ratio on utility. With a fixed number of applicants, total gain in utility is greatest when 50% of these men are to be accepted. That is to say, tests make the greatest contribution to the selection decision for which 50% of the a priori population can be utilized, other parameters being equal.

At extremely high or low selection ratios, gain in utility drops below zero, because the benefit from selecting superior men cannot compensate for the cost of testing. At these extremes the decision maker who tests is worse off than the one who fills his quota at random. Testing is unprofitable whenever y' becomes so extreme that $\xi(y') < C/\sigma_e r$; the location of this limit is shown in Figure 6. No matter what the selection ratio, testing is unprofitable if $C/\sigma_e r > .399$. In general, the limiting selection ratios are symmetric about .50; thus if $C/\sigma_e r = .05$, it is unprofitable to test if $\phi > .98$ or $< .02$.

Further insight into the role of selection ratio can be gained from a consideration of the average gain in utility per man accepted. This measure depends on $\underline{r}$ and σ_e in the same manner as total gain. Analysis of average gain shows that one may profitably vary the selection ratio if this is allowable. The relationship of average gain to ϕ , for a given $\underline{r}$, has been described by Brogden (11), and is shown in Figure 7 (see (1.10)). Average gain increases to a maximum as selection ratio decreases and then drops rapidly, falling to $-\infty$ as selection ratio approaches zero. The average ability of individuals accepted increases if selection is made more severe, but the proportionate cost of testing per man accepted increases. Varying the selection ratio (or cutting score) in a particular problem, we can arrive at an optimal strategy. The selection ratio which yields greatest average gain in utility depends on the ratio of $\underline{C}$ to $\sigma_e r$ (cf. (1.13)). This relationship is shown in Figure 6.

In many selection situations, the decision maker has a fixed quota to fill but may vary the number to be tested. The industrial tester, for example, may have a certain number of vacancies in a given period, but be able to alter

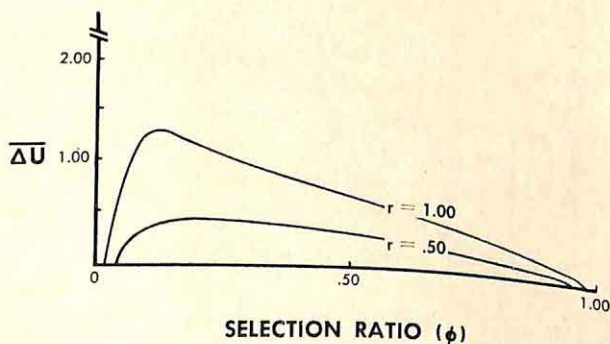


Figure 7. Average gain in utility per person accepted

the number of persons to be screened to fill these positions, reducing the selection ratio correspondingly. In this case the selection ratio of (1.13) can be interpreted as indicating the number of people to test relative to the number selected. The smaller the cost of testing, the larger the number of applicants it is profitable to test to fill the quota. It is always desirable (assuming fixed quotas and linear payoff) to test at least twice as many men as will be accepted if the test is worth using at all. At the low costs most likely to be encountered in practice, a large number of individuals should be tested relative to the number to be selected.

SELECTION WITH ADAPTIVE TREATMENT

It is reasonable to suppose, as Chapter III pointed out, that the decision maker will often be allowed to adjust the treatment according to the quality of the men he accepts. Such "adaptive treatment" may be expected to yield greater benefit than is obtained under fixed treatment, save where by good fortune the fixed treatment happens already to be fitted to the accepted men.

A general payoff surface

The first step, in undertaking a study of adaptive treatment, is to introduce a general payoff surface linking payoff functions for different treatments.

A postulated aptitude factor. When only one treatment and one test are under consideration, the contribution of testing can be studied adequately in terms of the relation of test score to criterion. The payoff function is the regression of the evaluated criterion on the test score. To deal with competing treatments, it will be convenient to postulate an aptitude factor $\underline{s}$.

The advantage of the intervening variable $\underline{s}$ is that it permits us to separate aspects of the decision problem chiefly associated with the differences between treatments from those chiefly associated with the test. Where several tests measure the same underlying variable, and we wish to consider their relative effectiveness, the introduction of $\underline{s}$ permits us to invoke the same payoff function with relation to all tests. Each treatment has its own payoff function; but if all the treatments under consideration depend on the same aptitude, it is much simpler to describe the relation of payoff to this aptitude than to have a separate payoff function for each test-treatment pair.

As explained in Chapter III, test scores may be factored, and among those factors the ones which account for variance in payoff under any of the treatments being considered are referred to as aptitude factors. That is to say, any factors which, when removed from the matrix of r_{ye_t} , reduce all residuals to zero constitute a set of $\underline{s}$ dimensions. There will not in general be a unique set of $\underline{s}$'s for a given body of data, since these factors may be rotated in various ways. One would ordinarily rotate in one of the many ways that would reduce the number of $\underline{s}$ dimensions to a minimum.

In studying selection and placement, we restrict ourselves to the case where only a single aptitude dimension is required to account for all communality between test scores and payoffs. If we are dealing with a single test, as we do in most subsequent chapters, the "true score" on the test may be regarded as such an $\underline{s}$ dimension. It is assumed that the test and payoff under any treatment have no common variance save that accounted for by $\underline{s}$, so that $r_{ye_t} = r_{ys} r_{se_t}$. We shall not extend our analysis of adaptive treatment to the classification problem, where multivariate information is employed and more than one $\underline{s}$ dimension is involved in decisions. The study of adaptive selection and placement requires assumptions going some distance beyond presently available data, and it appears wise to delay extension to more complicated cases.

The function relating expected payoff to aptitude $\underline{s}$ is

$$e_{st} = \sigma_{e_t} r_{se_t} s + e_{ot} \quad [3] \quad (2.1)$$

Admissible treatments. As indicated in Chapter III, treatments may conceivably vary continuously with respect to many dimensions. For a general theoretical survey it is convenient to assume that the decision maker can draw the final treatment from among such an infinite array of possibilities, rather than from a few alternatives. Since we assume that the payoff for any treatment is a linear function of $\underline{s}$, it is possible to classify all conceivable treatments with respect to only two parameters, the slope and the intercept of this function. Furthermore, one and only one treatment with any particular slope need be considered. This is possible because, of two treatments having the same slope, the one with the higher intercept is uniformly preferable at all levels of aptitude. Only this treatment is admissible. For convenience, we write m_t for the slope instead of $\sigma_{e_t} r_{se_t}$.

Equations relating e , s , and t . To relate $\underline{e}$ to $\underline{s}$ and $\underline{t}$, we employ the simplest mathematical function which covers a reasonable variety of non-trivial relationships. The surface so defined will fit a wide range of empirically obtained payoff functions if parameters are appropriately chosen. Appendix 2 states the specific assumptions leading to the following equation:

$$e_{st} = m_{st}s + c + b m_{st} - a m_{st}^2 \quad [4] \quad (2.3)$$

We may substitute $r_{ys}y$ for $\underline{s}$ where desirable. This surface, depicted in Figure 8, is a hyperbolic paraboloid with these properties:

1. There is a saddle point where $s = -b$. Through this point pass two straight lines that lie in the surface.
2. All horizontal cross-sections (equal payoff) are hyperbolas asymptotic to these lines.

3. All cross-sections where $\underline{t}$ is a constant are straight lines.
4. All cross-sections where $\underline{s}$ is constant are parabolas. This implies that there is a best treatment for each value of $\underline{s}$, and that the loss from assigning a person to treatment $\underline{t}$ rather than the t_s optimal for him is equal to $a(m_t - m_t_s)^2$.

The parameters $\underline{a}$, $\underline{b}$, and $\underline{c}$ depend upon the particular $\underline{s}$ linking treatments and tests.

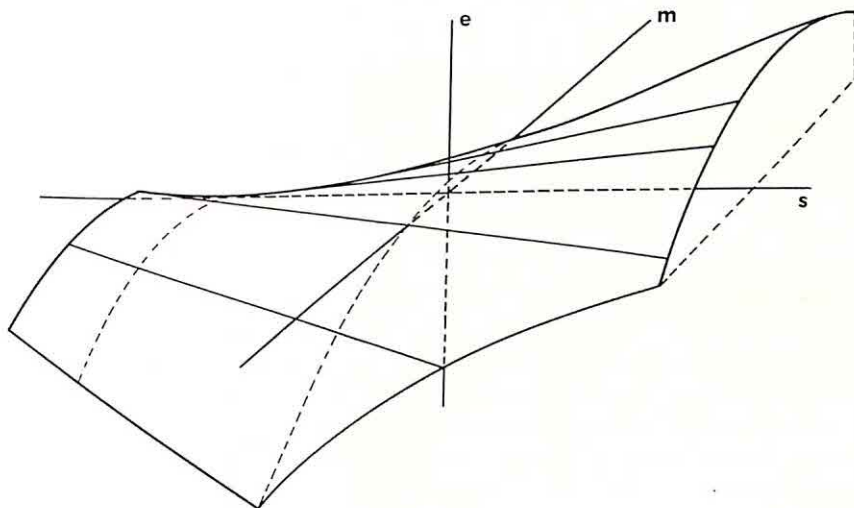


Figure 8. Expected payoff as a function of aptitude and treatment

Utility in adaptive selection

The formulas for estimating gain in utility are developed in Appendix 3. In evaluating the benefits gained from selection with adaptive treatment, and contrasting them with fixed-treatment benefits, one must reconsider not only the relation of payoff to score, but also the entire definition of gain in utility. In fixed treatment, a priori utility and a posteriori utility are simply the benefits when selected and unselected men are given the same treatment. But now the a priori treatment and the a posteriori treatments can be selected by the decision maker.

The best a priori treatment is t_o , a treatment suited to the average unselected man. Assuming a quota $\phi(y')$, the a priori utility is

$$U_{ot_o} = \phi(y')(c + b^2/4a) \quad [5] \quad (3.2)$$

The parameters are those of the utility surface [4].

So long as we confine ourselves to tests measuring a single aptitude $\underline{s}$, the choice of treatment to be given after testing depends on r_{ys} . The more valid the test, the more the aptitude of the selected men will depart from the average and the farther the selected treatment t_s will depart from t_o . The a posteriori utility is

$$U_{yt_s} = \phi(y') \left[\frac{1}{4a} (r_{ys} \frac{\xi(y')}{\phi(y')} + b)^2 + c \right] - C_y \quad [6] \quad (3.13)$$

The striking feature of this equation is the introduction of an r^2 term in addition to a linear term. In effect, by raising the quality of the men and simultaneously changing the treatment to take full advantage of this quality, adaptive treatment "compounds" the gain from any improvement in test validity.

The benefit from testing

$$\Delta U = U_{yt_s} - U_{ot_o} = \frac{r_{ys}^2}{4a} \frac{\xi^2(y')}{\phi(y')} + \frac{br_{ys}}{2a} \xi(y') - C_y \quad [7] \quad (3.8)$$

With adaptive treatment, the gain from testing is a parabolic function of r_{ys} . We will arrive at a fuller understanding of why the gain from adaptive selection takes this form after examining adaptive placement in Chapter VI.

Equation [7] permits us to generalize only about the relation of utility to the validity of the test as a measure of aptitude $\underline{s}$. The usual validity coefficient r_{ye_t} is equal to $r_{ys}r_{se_t}$ under our assumption, and this value may be expected to vary from treatment to treatment even when only one aptitude is involved. Since the slope of the payoff function $m_t = \sigma_e r_{se_t}$ is different for each treatment, it follows that the value r_{se_t} will also in general be different so that there can be no single validity coefficient in an adaptive-treatment problem.

If two or more tests are under consideration which measure different aptitudes, their relative value in adaptive selection (or placement) is difficult to describe. For each aptitude there will be a different payoff surface, involving new values of a and b. All subsequent statements in this chapter referring to the relation of utility in adaptive selection to "validity" are restricted to the case where the aptitude linking test to treatments is the same for all tests compared.

Comparison with benefit from fixed treatment

To compare fixed with adaptive treatment, we rewrite equation [2] in this manner:

$$\Delta U = m_t r_{ys} \xi(y') - C_y \quad [2]$$

In order to examine a specific case, let us assume that the treatment originally

in use is that best suited to the average man, i.e., $m_{t_A} = b/2a$ (cf. (3.1)). Such a situation could arise if previous selection policy had been fairly satisfactory, but it is thought that greater benefit will result if men of even higher quality are obtained. Using the same treatment after testing and a fixed selection ratio, the gain in utility is

$$\Delta U = \frac{b}{2a} r_{ys} \xi(y') - C_y \quad [8]$$

If, however, the institution were to now change the treatment to that optimally suited to the average ability of the selected men, the gain in utility is increased by $\frac{1}{4a} r_{ys}^2 \frac{\xi^2(y')}{\phi(y')}$. This additional gain might be found, for example, when introducing tests in a school selection program where teaching methods are subsequently adjusted to the higher level of ability of the accepted students.

When the a priori treatment is not that best suited to the average man, comparison between fixed and adaptive selection becomes more complicated. One must fix in mind the levels of utility obtainable by the following strategies:

1. Random selection, fixed treatment t_A U_{ot_A}
2. Random selection, treatment t_O suited to average men U_{ot_O}
3. Random selection, treatment suited to aptitude level $\underline{s}$ U_{ot_s}
4. Selection by test, treatment t_A U_{yt_A}
5. Selection by test, treatment t_O U_{yt_O}
6. Selection by test, treatment t_s U_{yt_s}

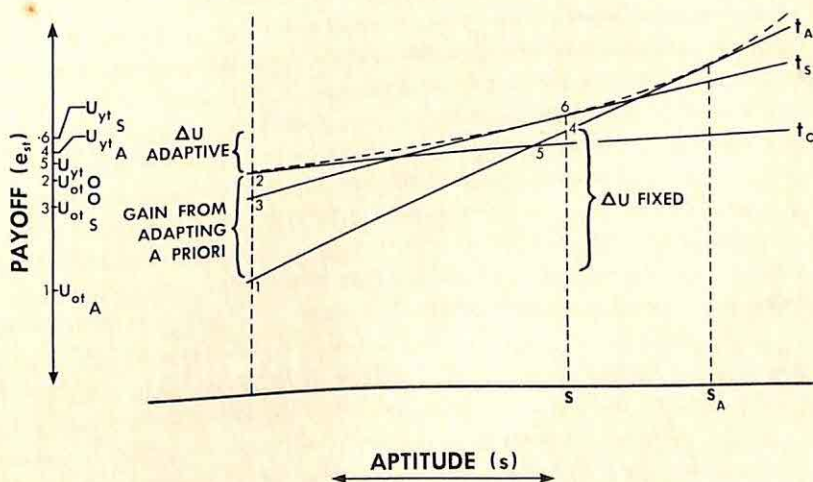


Figure 9. Comparison of gains from fixed and adaptive selection

For the sake of illustration (Figure 9) we may assume that the treatment t_A originally in use is that suited to very superior men whose average aptitude is s_A . Treatment t_o is the best of all treatments for a group of men whose average aptitude is zero. Finally, if the men actually selected have the aptitude s (in the figure, $0 < s < s_A$), the payoff function for the best a posteriori treatment is that marked t_s . In addition to the payoff lines for three treatments, the figure shows a curve to which all such functions are tangent (2.6). The tangent to the curve at any value of s shows the payoff to be obtained by giving the optimum treatment to men having that average level of ability. On this diagram, the six levels of utility are indicated.

The gain in utility with fixed-treatment t_A is represented by the difference between U_{ot_A} and U_{yt_A} (cost of testing here being disregarded). The final utility is, in this instance, distinctly less than may be attained with a less demanding treatment. If one were to change from t_A to t_o without selecting men, a benefit would accrue which is actually nearly as large as the benefit (ΔU fixed) reportedly attainable by testing. The gain due to testing should be considered as the gain beyond this best a priori utility, and is represented by ΔU adaptive ($U_{yt_s} - U_{ot_o}$). Other diagrams can be examined where the fixed treatment is located so that $s_A < s$, or even $s_A < 0$. The following generalizations apply to all such combinations.

Even though adaptation of treatment is common in practice, discussions of the value of testing in selection have overlooked this possibility, clinging to fixed-treatment assumptions. The utility from selected men is always greater with adaptive treatment than with fixed treatment, save at that one value where $s = s_A$. To a certain extent, therefore, these discussions have underrated the utility that can be obtained using the test. On the other hand, the gain due to testing with adaptive treatment is less than the gain under fixed treatment, because the a priori utility can be much higher under adaptive conditions. By including a component which could be achieved by altering the treatment to use unselected men more profitably, these discussions overrate the contribution of testing. Which effect is the larger would depend on the particular utility surface, the location of s_A , and also on the cost of adaptation. If practical conditions make it unreasonable to adapt treatments to men of average ability, then testing does have the value claimed for it in equation [2] where utility is linearly related to validity. But if treatment is "fixed" only because the question of adapting it has not been raised, the advantage claimed in the linear relation is much too great and the test is being credited with contributions which could be obtained in another manner, perhaps less expensive in the long run.

The amount the test adds to the best a priori utility is described by the surface shown in Figure 10. All cross-sections with selection ratio constant are parabolas. Since we are comparing tests involving the same aptitude, validity is here represented by r_{ys} , not r_{ye} . With r_{ys} constant, the cross-section is asymmetric, having a maximum at a selection ratio somewhat less than .50. The surface plotted is based on these parameters: $a = 1$, $b = 1/2$, $C = .05$.

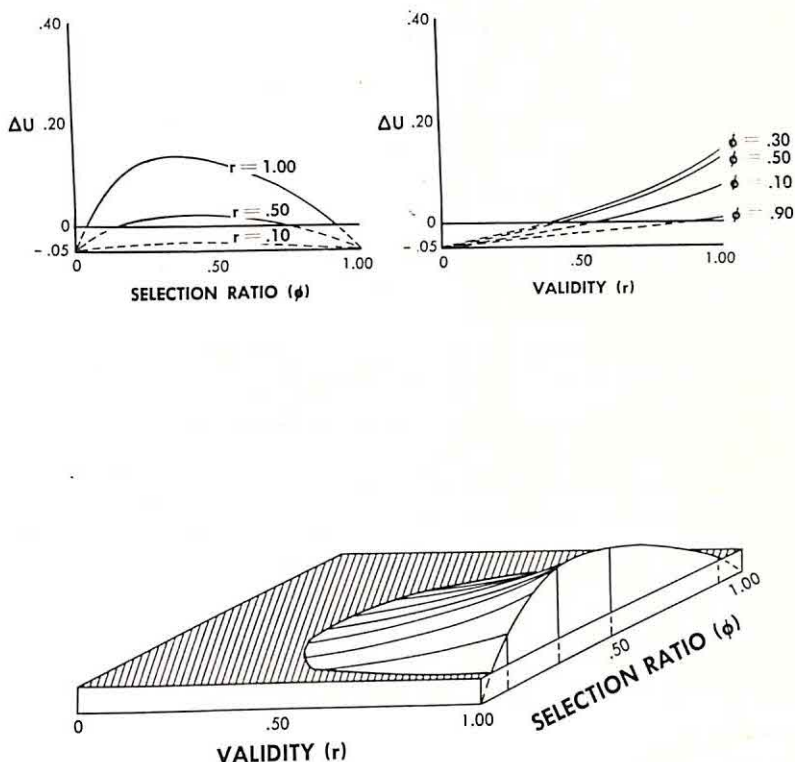


Figure 10. Relation of gain in utility to validity (r_{ys}) and selection ratio in adaptive-treatment selection

Implications for choice of tests

Persons validating selection tests should consider explicitly what might be achieved by adapting treatments. It is necessary to determine empirically the shape of the payoff surface wherever adaptation is allowable. Such investigations might well begin by comparing payoff functions for only two alternative treatments aimed toward the same goals, but ultimately it will be necessary to consider cases involving several continuous treatment dimensions.

The traditional sequence of industrial selection research has been to establish a job and associated training procedure, and then to seek predictors which will weed out applicants likely to perform poorly. It is appropriate to fix production goals without reference to the available selection tests, but there is no reason at all to regard the job organization or training methods used to attain those goals as being independent of such predictors. If payoff functions differ when different training methods are used, and these functions intersect somewhere in the range of ability, then it may be possible to modify the training with advantage. Determining objectives is the first step in selection research; after that, the research should seek to identify the best combination of tests, training method, and job organization.

Testing can bring about two types of change: an increase in the average quality of accepted men; and a further increase in benefits when treatment is adapted to fit this new level of quality. When treatment can be modified, it is possible that much of this increase in utility can be attained by adapting treatment to fit the level of ability of unselected men. The relative advantages of selection without adaptation, adaptation without selection, and adaptation with selection depend on the parameters of the postulated payoff surface (and on the cost of adaptation). No unqualified generalizations as to proper practice can be made without considering particular surfaces established empirically.

In a sense, adaptation permits one to trade increased training cost for attrition in screening. If one insists on utilizing only the cream of the applicants, the training can be less costly or can attain a higher terminal level. If a lower grade of applicant is accepted, one takes best advantage of his ability with a longer and less demanding training program. The quality of men selected depends on many things: on the social conditions which make the applicant pool larger or smaller, on the personnel requirements of other programs, and on the validity of the relevant tests. If two aptitudes are relevant to the same job, adapting treatment to fit men near the average on one aptitude would permit more severe screening on the other. The optimal strategy would be a complex mixture of adaptation and selection.

Recognizing the possibility of adaptive treatment provides a new way of looking at virtually every problem of test utilization. It seems likely that new results will emerge whenever this hitherto suppressed consideration is brought into test theory.

THE INTERPRETATION OF VALIDITY COEFFICIENTS IN SELECTION

The preceding pages deny the possibility of any simple answer to the query, "How valuable for selection decisions is a test with validity r ?" Neither the index of forecasting efficiency $\bar{E}$ nor the coefficient of determination r^2 is directly related to utility when payoff functions are linear. Neither, therefore, seems appropriate as an index of selection efficiency. Moreover, our utility functions bear only a limited resemblance to those of Taylor and Russell. And while utility is linearly related to validity in fixed-treatment selection, we find that the linear relation does not apply in selection with adaptive treatment. We may now integrate the various approaches to the interpretation of validity coefficients in selection, and point to some differences in the various rationales.

The linear relation

Brogden's linear relation between validity and utility describes the usefulness of a test (or composite score) under the following conditions:

1. Persons are divided into an accepted and rejected group. The rejected group is eliminated from the institution, while persons in the selected group are given uniform treatment.

2. The proportion of persons to be accepted is specified.
3. The treatment to be applied to accepted men is fixed a priori.
4. Payoff from accepted persons has a linear regression on test score.

The utility of a decision to reject is the same for each individual.

This implies that decisions are evaluated in terms of institutional utilities.

The Taylor-Russell results

The relation between validity and utility reported by Taylor and Russell is rather similar to the linear relation under many conditions. The chief differences are that (a) they report different shapes for the validity-utility relation at different selection ratios, and (b) the relation becomes decidedly non-linear at high validity or when a high proportion of unselected men are judged successful.

The Taylor-Russell approach assumes fixed treatment and is otherwise like the Brogden analysis, except that outcomes are evaluated differently. Taylor and Russell ignore rejected men, in effect assuming their contribution to be zero. Accepted men are classified into "successful" and "unsuccessful" groups, and all men in each group are regarded as making equal contributions. Thus, while Brogden assumes a continuous equal-interval scale for payoff at different levels of criterion performance, Taylor and Russell assume a discontinuous two-valued payoff. The two scales are related in such a manner as this:

c	20	25	30	35	40	45	50
e_c (linear assumption)	-1	0	1	2	3	4	5
e_c (Taylor-Russell)	0	0	0	0	1	1	1

The Taylor-Russell approach resembles that used in acceptance testing of industrial products. In that field, inspection plans are described by various tables indicating the efficiency with which each plan protects the purchaser against defective objects. The Taylor-Russell tables indicate the "average outgoing quality (AOQ)" for a personnel inspection plan, and are therefore comparable to the AOQ tables for industrial inspection. Each table indicates what proportion of "defective individuals" will be passed by a particular selection ratio, knowing the frequency of such individuals in the population.

Where unsuccessful men are discharged or removed from their assignments or fail to complete the training, the difference in utility between a successful and an unsuccessful man is likely to be great. Training effort is wasted; morale of all employees suffers; management time is consumed in making the decision to discharge. Such costs may be far more significant than the differences among the workers retained. Under some conditions, moreover, differences in ability beyond the minimum needed to perform the job do not lead to differences in benefit. If production is standardized either by job definition or by voluntary restriction of output, then benefits will depend little on ability. In these instances, evaluation by the proportion of successful workers among those accepted may be fully appropriate.

Sometimes differences in output are believed to occur, but are presently unmeasurable. Thus teachers who fail as instructors or class managers are discharged, but no agreement can be obtained on relative evaluation of the teachers retained. Likewise, in evaluating psychiatric screening a person who breaks down must be counted as markedly different in value from those who do not, but it is difficult to sustain any quantitative comparison of contribution among those who remain above the level of overt disturbance. This situation too justifies a step-wise evaluation.

Examination of the number of "hits" is sometimes accompanied by an examination of the "misses", or number of persons rejected who were of acceptable quality. The best-known discussion of this approach is Berkson's "cost-utility" paper (5). He defines "utility" as the proportion of unsuccessful men eliminated, and "cost" as the proportion of successful men rejected. In any selection scheme, raising the cutting score increases the Berkson "utility" with a corresponding increase in "cost". The decision maker seeks a strategy which balances these two risks most satisfactorily. A group of sociologists (28) has described several distinct formulas for evaluating hits and misses, which need not be reviewed here. Our thinking is most consistent with the plan which assigns particular values to "hits" and "misses", and adjusts the cutting score to maximize expected utility. In institutional selection with fixed quota where men rejected leave the institution, it is not meaningful to consider misses, however. Rejection of men of good quality does not decrease the output of the institution; their frequency bears only on "what might have been". Two-sided evaluations are appropriate only when the screening test divides men into two groups who remain within the institution but are treated differently. An example is the use of a test to predict parole, where the parolee and the man held in prison both affect the "balance sheet" of the correctional system. These, however, are placement rather than selection decisions.

Parabolic relations

Like others who have investigated the selection problem, we find no basis for interpreting the usefulness of tests in terms of the coefficient of forecasting efficiency or the coefficient of determination. These indices should not be employed in evaluating tests for selection purposes. Chapter VI will consider the appropriateness of these indices for certain other decisions.

Paradoxically, having dismissed two traditional parabolic functions, we find that another parabolic relationship is appropriate for evaluating tests in selection. Where adaptive treatment is allowed and all tests being compared measure the same aptitude, equation [7] -- or some similar function based on a payoff surface of degree higher than two -- is the most adequate statement of the relation of utility to validity. When the selection ratio is moderate to high, this

parabola has little curvature and the linear relation approximates it fairly well. At low selection ratios, concave curvature may be appreciable (quite contrary to the Taylor-Russell picture of convex curvature at low selection ratio).

By neglecting the possibility of adaptation, some discussions of the linear relation between utility and validity tend to give a misleading picture of the value of testing per se. If the treatment presently in use is fixed, as it is in some industrial situations, small increases in validity of selection procedures will result in considerable benefit. If adaptation of the treatment is possible, however, and selection ratio is low, much of the gain can be achieved by adjusting the treatment to the average man and thus raising the utility of the a priori strategy. If selection is then employed, the increase in utility which results is linearly related to validity, but it is possibly quite modest in amount. Further adaptation of treatment to the selected group brings additional gains which are parabolically related to validity.

Our analysis of adaptive selection surely calls for a reconsideration by those who have used the linear relation to "prove to management" that tests are beneficial. The "profit" they have piled on the scales includes a substantial bag of gold that might be earned by another branch of the business if testing were abolished. Testing is usually beneficial under adaptive conditions, but not as beneficial as has been claimed when only fixed treatment is considered.

"What is a test of given validity worth?" is a complicated question. The total benefit achieved through use of a test is of more concern to the institutional user than is the proportion of possible benefit or improvement obtained. This total benefit is increased, generally speaking, with increases in the number of persons to be tested, the importance of the criterion performance to the institution, and the extent of individual differences in performance. The benefit decreases with greater cost of testing. Therefore a test of validity .20 in one situation may be more beneficial than a test of validity .60 in another. The characteristics of the specific decision determine "what the test is worth".

5

TWO-STAGE SEQUENTIAL SELECTION

The preceding chapter assumed that the selector would make a single terminal decision. The use of investigatory decisions, mentioned in Chapter II, will now be considered. Efficiency of testing is often improved by a sequential plan allowing the decision maker to continue testing whenever he is in doubt about acceptance or rejection of an individual.

POSSIBLE USES OF SEQUENTIAL METHODS

Sequential acceptance testing

Sequential methods were first developed to meet the requirements of industrial inspection, but their scope has been extended to cover all testing of statistical hypotheses. The quotation from Girshick (see page 2) indicates that they are now "the rule rather than the exception" in statistical decisions. The original industrial inspection problem is remarkably similar to personnel selection. Either the manufacturer or the consumer inspects samples of products to determine whether a given lot is of acceptable quality. This is precisely analogous to deciding whether an applicant is of acceptable quality on the basis of a sample of his behavior, and one may therefore anticipate gains from transferring the inspection methods to personnel work.

Sequential plans are strategies in which investigatory decisions are allowed. In "double-sampling" or two-stage plans, a terminal decision is required after the second testing. Multi-stage plans are also possible, investigatory decisions being allowed after every stage; some terminal decisions are made after each test but a steadily decreasing fraction of the cases are carried through subsequent tests.

Sequential plans are beneficial because it costs something to gather information. Obviously, it would always be better to administer the full series of tests, whatever its length, to every person if observations cost nothing. As was pointed out in the preceding chapter, test theory has generally ignored costs of

testing, but even quite small costs affect choice of testing procedure. We may anticipate economies from adoption of a sequential plan in personnel selection.

Considering only cost of testing, maximum benefit would always be attained by using a multi-stage plan. Administering a sequential selection plan, however, involves costs which would not be found in single-stage testing. A sequential plan requires intelligent administration; in industry, the fact that varying lengths of time may be required to inspect different objects sometimes creates serious scheduling problems. Likewise, in individual processing of recruits or applicants, the smooth flow of a processing line would be disrupted if time for one of the examinations varied substantially from person to person. Where a complete sequential plan would be awkward to administer, a double-sampling plan is frequently a satisfactory compromise. It retains much of the efficiency of the multi-stage plan, and yet is relatively easy to administer.

The present chapter explores the characteristics of two-stage sequential sampling plans for personnel selection. This analysis will not *per se* indicate just when sequential plans should be used, since the cost of administering the plan is not included in the mathematical specifications. (In this, we follow the practice of statisticians.) While our analysis will show benefits from the adoption of sequential procedures, these benefits must always be weighed against the unspecified administrative costs.

Experience in other fields suggests that sequential procedures attain a given level of accuracy of decisions with about half the amount of testing required by a single-stage plan. Arbous and Sichel (3) have studied a "pre-screening" procedure for personnel testing which is in effect a modified double sampling. Although their procedure is not the optimum sequential plan, they report noteworthy savings in testing time. A similar pre-screening plan was developed independently by Cochran (20). He points out that in a great variety of situations it is advantageous to reduce the original group by stages. He gives as one example the selecting of hogs for breeding, where some of the original group can be dropped because of inferior weight. Final selection among those who pass the weight test is based on the quality of their first offspring (a test which is expensive to make because it requires long delay). High cost of testing, as when costly radar equipment is used in a proficiency test (cf. Evans (33)), makes sequential methods especially beneficial. Particular advantages from sequential testing are also to be anticipated when men are tested individually, since cost of testing is then relatively great.

Other sequential decisions

Our interest in sequential methods is not confined to a literal application of the acceptance-testing methods of industry. Sequential approaches to many other problems may be equally important, as Fiske and Jones (34) demonstrated by their review of psychological applications.

The sequential method departs from the traditional psychometric procedure of standardizing test questions so that all persons are tested in the same

way. Questioning is pursued only until the required decisions can be made; testing time can therefore be used more efficiently. Many forces appear to be compelling increased concern for efficiency in testing. For one thing, the identification of more and more aptitudes relevant to various occupations or job assignments means that thorough testing over all significant fields will henceforth require more time than is usually available. In proficiency and achievement testing also, it is increasingly recognized that a very large number of outcomes need to be measured in order to assess the individual and his instruction. A further reason for highly efficient testing is that tests are being introduced where only very small amounts of testing time are available. This is notable in military research where it is desired to test men on active duty in advanced areas. Such pressures already confront psychologists with demands to produce a test to measure this or that ability in five minutes of testing time. Test designs therefore must be maximally efficient.

Frequently the tester is required to decide among several treatments for each individual. Job classification is an example. In clinical testing also, one ordinarily has many hypotheses about the individual which he may accept or reject. In these instances, the problem at the first stage of testing is to divide the alternatives (rather than the people) into three classes: those accepted, those rejected, and those to be studied further. What alternatives remain to be considered will determine the character and length of the subsequent tests. Present procedures in guidance and clinical testing are obviously sequential in nature; this provides considerable motivation for developing adequate theory regarding sequential testing. Such a theory not only can clarify what is now going on, but should ultimately lead to substantially better procedures for "sequential analysis within the individual".

DETERMINATION OF OPTIMUM STRATEGY

The two-stage selection procedure involves two tests y_a and y_b , which may be correlated. One test is administered to every man; the second test is given only if the decision-maker wishes further information before making a final disposition of the individual. The first test may be one of a group of dissimilar tests or it may be a short test made up of representative items from the complete battery, the remainder of the battery being considered as the second test. As a matter of convenience, we express the scores on the two tests in terms of independent components, y_1 being the standardized score on the first test and y_2 being the standardized score on the other test after the first test is partialled out.

Alternative testing plans

At least five different selection strategies are available (without considering the possibility of altering the lengths of tests y_a and y_b):

1. Non-sequential battery. Administer the total battery to all men, and accept those having highest scores on an optimally weighted composite. This is the conventional selection method, and we shall hereafter refer to it as the "Battery" procedure.
2. Single screen. Administer either test alone, and base all decisions on it.
3. Sequential. After the first test is given, divide the group into three portions: those accepted, those rejected, and those to be given the second test. Base final decisions for the last-named men on the composite of both tests. We hereafter refer to this as the "Sequential" strategy even though Pre-reject and Pre-accept are also sequential devices.
4. Pre-reject. After the first test, reject some men and continue to test all others. For these others, base final decisions on the composite of both tests.
5. Pre-accept. After the first test, accept some men and continue to test all others. For these others, base final decisions on the composite of both tests.

Figures 11 and 12 will aid in distinguishing these procedures. Figure 11 shows the joint distribution of y_1 and the independent second variable y_2 . Three lines are located in this figure. The line at y_1' is a cutoff below which

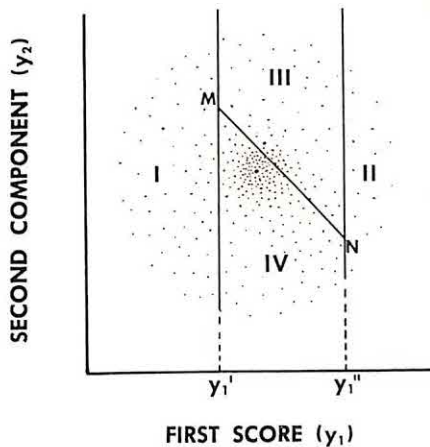


Figure 11. Sequential selection procedure

persons are rejected after the first test. The line at y_1'' is a cutoff above which persons are accepted on the basis of the first test alone. The slant line cuts off persons whose weighted composite score $Y = f(y_1, y_2)$ is above a certain level. Any person is accepted whose composite score on both tests is above

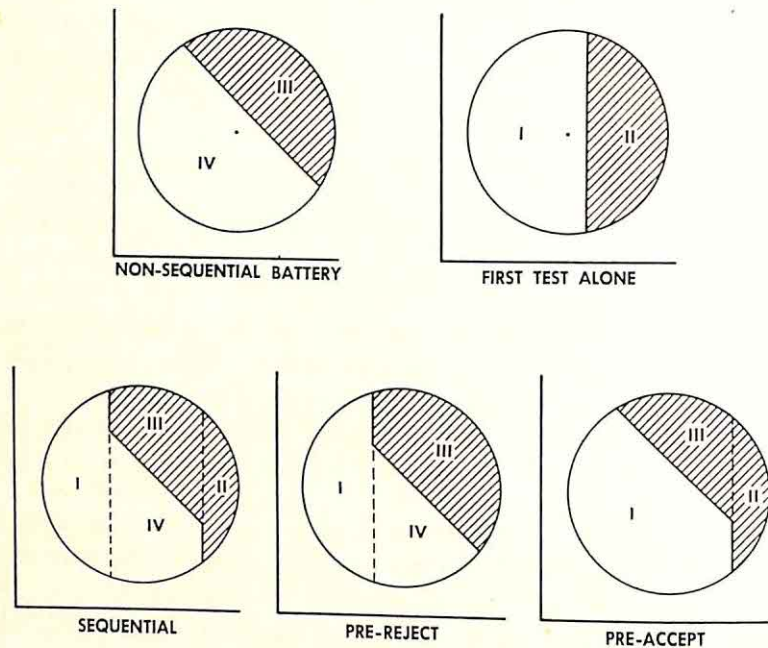


Figure 12. Alternative methods of selection

the line, and vice versa. The lines divide the distribution into areas I, II, III, and IV. Each strategy represents a different treatment of these areas, as shown in Figure 12.

The Pre-reject strategy is a version of the Arbous-Sichel method. They propose that an inexpensive portion of the battery be used as a first screen to eliminate those men who have very low probability of passing on the battery as a whole. Their method is not optimum, since it neglects the possibility of accepting some men on the first screen and does not allow for adjusting risks according to the cost of the second test. Cochran's method also is a pre-reject procedure. While he argues for the importance of taking cost of testing and selection ratio formally into account, his procedure does not do so. Indeed, he says (20, p. 455) that there "does not seem to be a useful general solution in functional terms". Our result in Appendix 4, however, is offered as such a solution for those situations where a linear payoff function applies.

Evaluation assumptions

At least three different methods of specifying evaluations are available, any one of which leads to a somewhat different location of the cutoff points y_1' and y_1'' . The first is the use of a continuous scale for evaluating outcomes in utility units and specifying a relation between payoff and test information. If

the a priori distribution of test scores and payoffs is known, and if fixed treatment is used, the strategy which maximizes payoff can be determined for a given cost of testing. This method is employed in Appendix 4, using a linear relation of payoff to test score and assuming a normal distribution of scores and payoffs. While such a system of cardinal utilities is mathematically satisfying, it requires difficult computations. Approximate procedures based on Neyman-Pearson risks (see below) are therefore generally used at present to study practical problems. We shall use such a method to study multi-stage sequential procedures in Chapter VII and Appendix 6.

Neyman-Pearson risks were introduced into general hypothesis testing as a modification of the confidence-level system of Fisher. While Fisherian analysis reports the risk of rejecting a null (or other a priori) hypothesis when it is true, the investigator wishes also to consider the risk of accepting this hypothesis when some alternative hypothesis is true. Neyman and Pearson drew attention to these two kinds of "errors", and used them to compare the power of various statistical tests. A decision maker is required to indicate the errors he wishes to avoid and to state what risk of each error he is willing to tolerate. (Unless he can tolerate some degree of error, he cannot use a sampling plan to get information.) In a selection plan, the decision maker might wish to eliminate men whose true aptitude is below a standard score of $+5$, for example. It would not ordinarily be a matter of concern if the plan admitted some men at $.48$ or rejected some at $.52$, but there is some magnitude of error which would be serious. The decision maker is perhaps willing to admit some men as low as $.35$, on the theory that they can be discovered and eliminated, and wishes to minimize the risk of missing men of aptitude $.55$. Thus we have two tolerance limits, $.35$ and $.55$. He now specifies the risk for each error, perhaps stating that he is willing to tolerate admitting 10% of the men at aptitude $.35$ (the risk will be less than $.10$ for men below $.35$), and will tolerate a risk of missing 2% of the men at $.55$ (less above that point). In the language generally used for this system, the α risk is said to be $.10$, and the β risk, $.02$. Once these are specified, the selection procedure requiring the minimum amount of testing can be readily determined. This strategy is not dependent on an assumed a priori distribution. To compute the expected net utility it would be necessary to make such an assumption, however, since different distributions result in a different total frequency of each kind of error.

A selection plan based on Neyman-Pearson risks does not weight costs of testing explicitly. Instead, the decision maker is required to take them into account implicitly in setting tolerances and risks. Determination of a sequential plan and study of its performance characteristics has been reduced to a simple routine procedure (35, 61) by studies performed during World War II by the Statistical Research Group at Columbia University, working under the leadership of Wald. It is generally believed that the methods based on risks are an adequate approximation to the procedures based on a cardinal utility scale, for those cases to which they apply.

The evaluation of strategies in terms of the number of "hits" and "misses", discussed above in connection with the work of Taylor-Russell and Berkson, might also be applied to sequential problems. Indeed, Arbous and Sichel employ this evaluation to describe the efficiency of their plan even though they employ risks in developing it.

Location of cutoffs

In our treatment of the double-stage sequential procedure, the cutoff scores on the first test and on the total battery (the lines y_1' , y_1'' , and MN in Figure 11) are to be located so as to maximize utility for any selection ratio and strategy, using cardinal utilities and costs. The mathematical solution is presented in Appendix 4. The same cutoff y_1' is associated with a particular MN for both Sequential and Pre-reject strategies, and the same y_1'' for Sequential and Pre-accept. Whichever test is designated y_1 will be given to all men. Persons falling in any score array on y_1 are accepted or rejected at once if screening on the second test will not raise their total contribution enough to compensate for the cost of testing.

There is no simple formula to locate y_1' and y_1'' directly, but the vertical coordinates y_2' and y_2'' of points M and N depend on only two parameters, C_{y_2} and r_{y_2Y} . These represent the cost of the second screen and its correlation with the total battery (remembering that any component correlated with the first test has been removed from y_2). The following relation specifies y_2' :

$$y_2' \phi(y_2') = \xi(y_2') - C_{y_2} / \sigma_e r_{ye} r_{y_2Y} \quad [9] \quad (4.18)$$

y_2'' is equal to y_2' and opposite in sign.

The foregoing equation is very much like equation (1.13), which expresses the optimum selection ratio in single-stage selection (Figure 6). An analogous cost-benefit ratio dictates the selection ratio within the array for the double-stage problem. For the lower boundary the second test should not be used to select among men in any y_1 array if the battery cutting score would accept fewer persons in the array than the optimum selection ratio for a test used singly which has cost C_{y_2} and validity equal to the independent validity of the second test. For the upper boundary the second test should not be used if the battery cutting score rejects fewer persons than the optimum selection ratio for maximum gain per man rejected when the second test is used singly.

In order to specify the sampling plan, we choose a cutoff value $Y = Y'$, where Y represents the best weighted composite of y_1 and y_2 . This cutoff leads to a different proportion of persons accepted under each of the strategies. By the method described in Appendix 4, y_1' and y_1'' are computed. The utility for any strategy is a function of the expected payoff of the men selected (summed over the shaded areas) less the cost of testing. The cost of testing is C_{y_1} ,

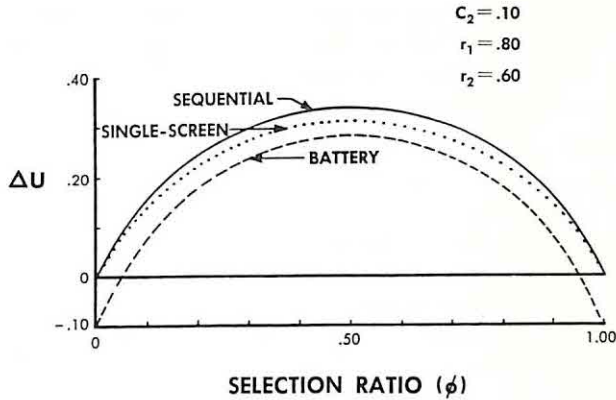


Figure 13. Gain in utility with sequential and non-sequential strategies

plus C_{y_2} times the proportion of persons given the second test. From (4.8) one can calculate the utility for any strategy and, from (4.9), the selection ratio corresponding to Y' . By taking a series of values for Y' , we generate a curve showing utility as a function of selection ratio.

Comparison of sequential and non-sequential strategies

Curves for these and other various sets of parameters were obtained by programming* the equations of Appendix 4 for the Illinois Automatic Electronic Computer (ILLIAC). Figures 13, 14, and 15 are based on three sets of parameters, chosen to illustrate several conclusions. An examination of Figure 13 will acquaint the reader with the type of information to be obtained. Utility per man tested is plotted as a function of selection ratio for three testing strategies. The figure is based on two parameters, r_1 (written $r_{y_1 Y}$ in Appendix 4) and C_2 (elsewhere written $C_{y_2}/\sigma_e r_{Y_e}$). Fixing r_1 determines the contribution of the second test to the battery, since the sum of squares of the independent contributions must be 1.00. Dividing cost by $\sigma_e r_{Y_e}$ (and also dividing ΔU by $\sigma_e r_{Y_e}$ before plotting) has no effect on the shape of the functions to be discussed and makes it simpler to compare various strategies since it reduces the number of parameters to be considered. In a practical problem the validity r_{Y_e} of the total battery is fixed by the choice of tests.

The cost C_1 of the first screen does not enter into the comparison of strategies per se, since it is charged against utility equally in all strategies. Thus for the purpose of comparing various strategies after a first test has been designated, no generality is lost by setting C_1 at zero. In general the cheaper test should be used as the first screen when the zero-order validities of the

*With the assistance of Jack C. Merwin

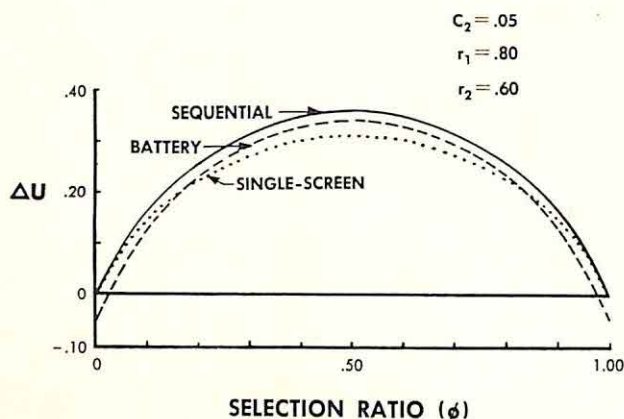


Figure 14. Gain in utility with sequential and non-sequential strategies

two tests are roughly equal or when the cheaper test is also the more valid. But if the more expensive test is the more valid, the choice of first screen depends on the selection ratio. In our analysis we tacitly assume that the test chosen as first screen is preferable at all selection ratios. In Chapter VII we consider the problem of adjusting the length of the first test optimally when it is a representative sample of the total battery.

The dotted line in Figure 13 indicates the utility obtained by administering the first test alone (Single-screen). This curve is similar to the cross-sections in Figure 5. The broken line indicating utility under the Battery strategy is likewise consistent with Chapter IV. In this figure, the cost of the second test is so great that Battery is never as profitable as Single-screen.

The solid line describes the utility yielded by the Sequential procedure at each selection ratio. At extreme selection ratios, Sequential comes close to

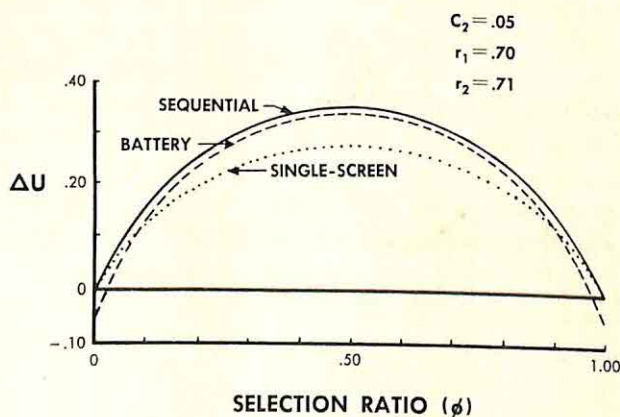


Figure 15. Gain in utility with sequential and non-sequential strategies

the curve for Single-screen. At all selection ratios, Sequential is equal or superior to either of the other strategies. We shall find that this is a general conclusion.

Figure 14 employs tests with the same relative validity as Figure 13, but the cost of the second test is lowered. This raises the utility for Battery at all selection ratios so that it is now superior to Single-screen at intermediate selection ratios. Sequential also yields more benefit than formerly, but its advantage over Battery has declined. Figure 15 retains the same cost, but raises the contribution of the second test and correspondingly decreases the contribution of the first. Comparing this to Figure 14, we find that the change has lowered the value of Single-screen, giving Battery greater superiority. But Sequential, which uses the second test only in relatively doubtful decisions, is still superior to both the other strategies.

In the subsequent illustrations pointing out various trends in the computed results, the parameters of the decision problem are generally taken to be $\phi = .50$, $r_1 = .70$, $r_2 = .71$ (which values would be equal save for the necessity of rounding in computation) and $C_2 = .10$. The values ΔU and C_2 are defined as in Figures 13-15. It should be noted that the vertical scale in Figure 16 is different from that of the previous figures.

The advantage of Sequential over Battery and over Single-screen is shown in Figure 16. At extreme selection ratios the advantage over the conventional Battery procedure equals the cost of the second test. At selection ratios near .50, Sequential has relatively little superiority. The figure emphasizes the comparison of the Sequential strategy to the better of the two non-sequential procedures. At extreme selection ratios it is compared to Single-screen, and at intermediate selection ratios, to Battery. As expected, Sequential is consistently superior by some slight amount, but its most striking advantage occurs neither at extreme selection ratios nor where $\phi = .50$. Battery and Single-screen yield equal utility at two selection ratios where cost and contribution

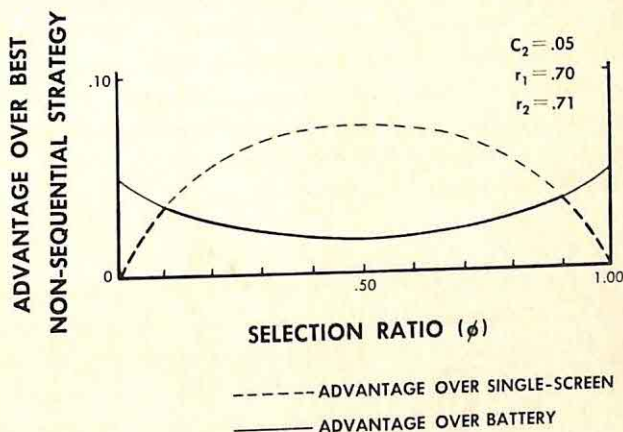


Figure 16. Comparison of sequential and best non-sequential strategy

of the second test are exactly balanced. Sequential always offers its maximum advantage at these selection ratios.

Effect of variation in cost

The cost of the second test obviously determines which of the strategies should be used. When a person is given the second test more information is obtained, but if cost of testing is high, the decision maker should seek this information only for the most doubtful decisions if at all. The change of utility with cost is shown in Figure 17, with r_1 and selection ratio held constant.

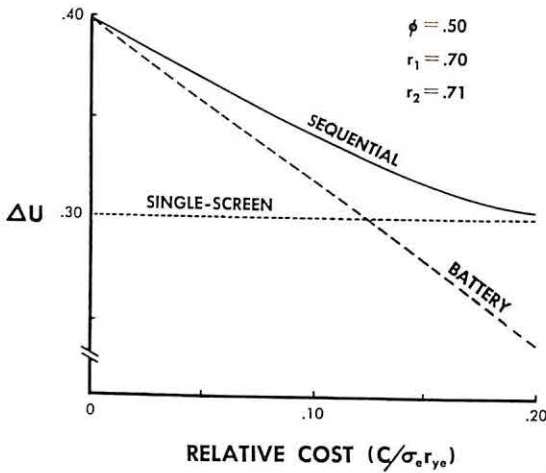


Figure 17. Utility as a function of cost of second test

The following points should be noted. As cost increases (or $\sigma_e r_{ye}$ decreases), Battery becomes less and less beneficial. At some point Single-screen becomes superior to Battery. At low cost, Sequential is slightly better than Battery; at high cost, slightly better than Single-screen. The greatest difference occurs, as before, where the Single-screen and Battery lines cross. The pictured relationships are typical, although the magnitude of differences depends on the parameters. It should be realized that if C_2 becomes very low, the test now given second would be the better first screen.

Effect of variation in r_1

The greater the weight of y_1 in the composite the greater the benefit from both Single-screen and Sequential (Figure 18). At low values of r_1 , Battery is about as good as Sequential. (Here it would perhaps pay to give the second test only.) As r_1 increases, the advantage of Sequential over Battery increases, but its advantage over Single-screen diminishes and finally becomes negligible. More extreme selection ratios or higher costs of second test would enlarge the region of r_1 in which Sequential is only slightly better than Single-screen.

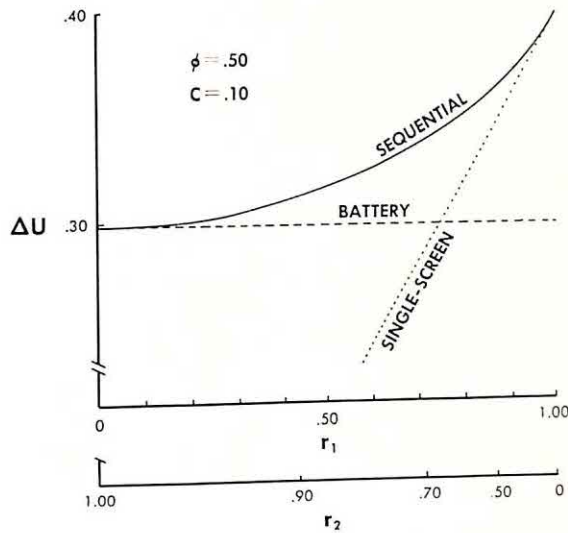


Figure 18. Utility as a function of relative validity of tests

PRE-REJECT AND PRE-ACCEPT STRATEGIES

Under all conditions of testing, the utility from the Pre-reject strategy lies between that of Battery and Sequential (Figure 19). When selection ratio is high, very few people will be rejected at the first stage and Pre-reject has negligible advantage over Battery; when selection ratio is low, Pre-reject is nearly as advantageous as Sequential. Pre-accept has a similar relation, approaching Sequential at high selection ratios and approaching Battery at low selection ratios. The curves for Pre-accept and Pre-reject are mirror images.

Improvement in utility for these incomplete sequential procedures, over the best non-sequential, is plotted in Figure 20 as a function of selection ratio.

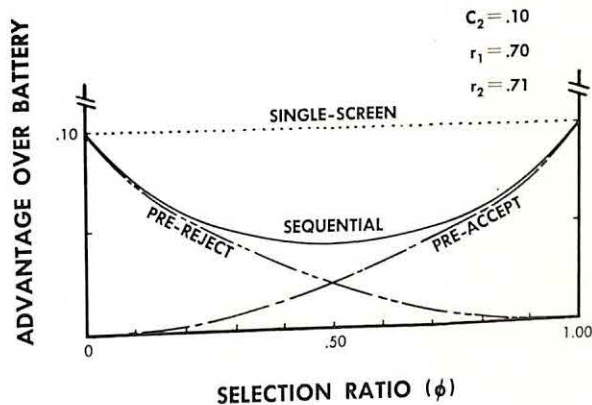


Figure 19. Utility with Pre-accept and Pre-reject strategies

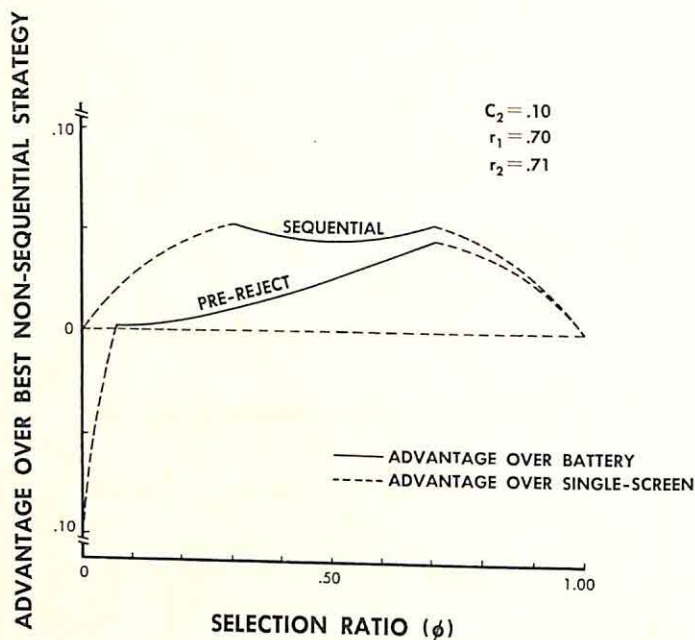


Figure 20. Comparison of Pre-reject to Sequential and best non-sequential strategy

Pre-reject gives only half the advantage of the full Sequential method when selection ratio is .50, and gives much less advantage at higher selection ratios. It is, however, as good as Sequential at low selection ratios.

We conclude that under our assumptions the proposal of Arbous and Sichel, and the complementary Pre-accept method are not superior to Sequential. There are, however, practical situations where a complete sequential strategy cannot be considered. For example, if the second "test" consists of admitting men to the training course and weeding out those who do poorly, "pre-acceptance" is meaningless since all men accepted have to go through the training. Cochran's breeding experiments are similar, in that second-screen data (quality of first litter) are necessarily obtained for all subjects not rejected.

AMOUNT OF TESTING SAVED

The core of the Sequential strategy is the correct determination of the cutting scores y_1' and y_1'' . Arbous and Sichel provide charts to indicate optimum cutting scores in their method, for various selection ratios and r_1 . In lieu of such charts or tables for our strategy, Appendix 4 outlines a procedure by which the tester may readily compute the desired cutoffs, knowing the parameters of his decision problem.

Typical results are plotted in Figures 21 and 22. In each diagram, the cutting scores are shown as a function of r_1 , assuming a cost of .10 units for the second test. Setting $Y' = 0$ (Figure 21) will yield a selection ratio of .50 regardless of the value of r_1 . When the first test accounts for nearly all the

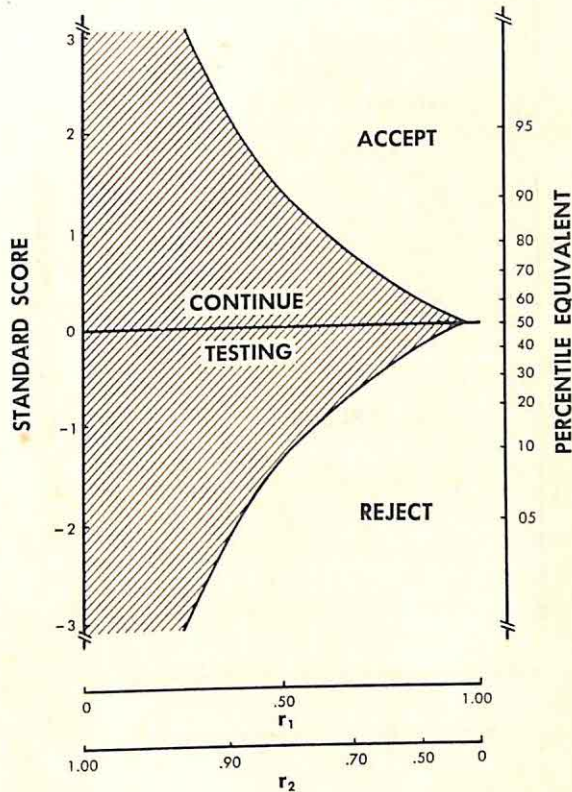


Figure 21. Cutting scores on first test when $Y' = 0$

battery variance, the two cutting scores coincide at zero, and no one is given the second test. As the contribution of the second screen increases, more and more persons are given the second test. When $r_1 = .80$ and $r_2 = .60$, the second test is given to persons with y_1 between $-.46$ and $.46$, i.e., to about 35% of the applicants. The Sequential strategy saves 65% of the second-screen testing ordinarily required.

One would be interested in the relation of first-screen cutoffs to r_1 for selection ratios other than .50, but to determine this relation would require a very large number of computations. Any given Y' yields a different selection ratio depending on r_1 ; to obtain a desired selection ratio, therefore, one must choose a value of Y' depending on r_1 . It is not too difficult to compute

first-screen cutoffs for a particular value of Y' . This has been done for $Y' = 1$; the results are plotted in Figure 22. When $r_1 = 1.00$ a selection ratio of .16 is obtained, but as r_1 decreases the selection ratio very rapidly decreases.

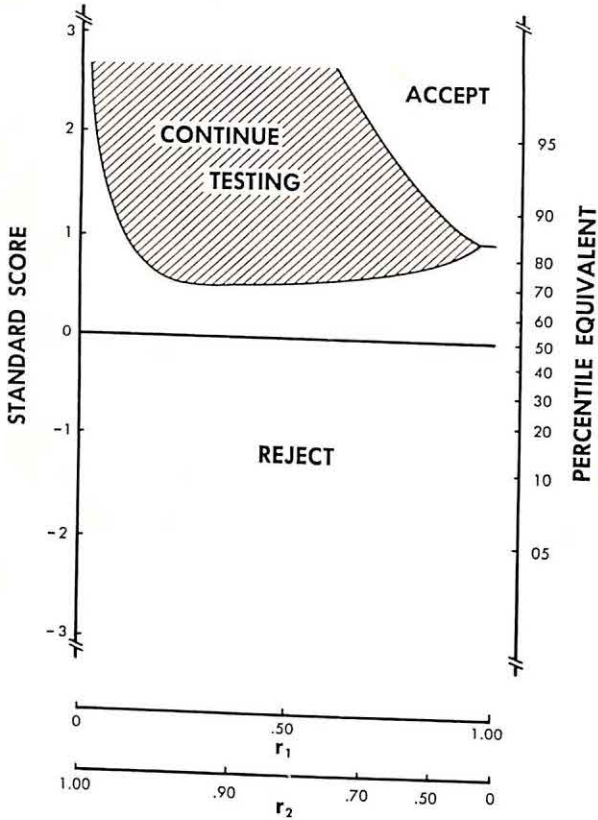


Figure 22. Cutting scores on first test when $Y' = 1$

One can see from Figure 22 that as the contribution of the second screen increases the proportion of persons given the second test increases and then decreases. This is not particularly meaningful, however, since when the battery score is determined almost entirely by the second test the selection ratio approaches zero for this value of Y' .

Figure 23 shows the proportion of the total group given the second test at each selection ratio, for one set of parameters. Also plotted is the proportion who would be given the test if Pre-reject strategy were used. A sequential procedure greatly reduces the amount of testing when the selection ratio moves toward either extreme. The striking advantage of Sequential over Pre-reject in terms of testing effort saved confirms our earlier conclusion.

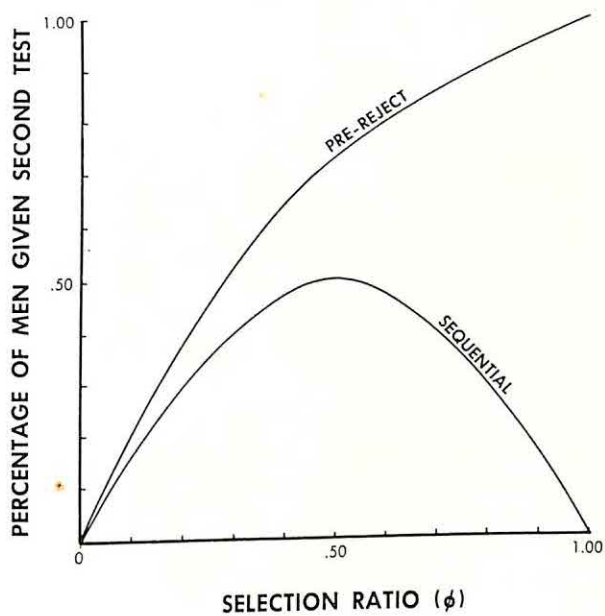


Figure 23. Proportion of persons given second test

6

PLACEMENT DECISIONS

In placement, persons are divided among two or more groups who remain within the institution and receive different treatments. Division of a group between two treatments has often been spoken of as screening or selection (e.g., "selection of students for an accelerated program", "screening of emotionally disturbed cases for individual examination", "selection of prisoners for parole", etc.). In each of these instances, the "selected" group is given one treatment while the others remain within the institution (broadly defined) and are given some other treatment. Therefore, these are placement decisions. Although placement may be considered as a special case of classification, it is of importance in itself because most measurement and prediction by means of tests can be interpreted in terms of the placement model.

In fixed-treatment placement with a limited number of categories, the score scale is divided into segments, and persons in each successive segment of the scale are assigned to a different treatment. Sectioning of students into classes according to learning ability is a prototype placement problem. Illustrations are also found in some types of job assignment, for example when a proficiency test is used to determine the level of responsibility (pay grade) to which a typist will be assigned. Each level has its corresponding duties and pay, and is characterized by a different function relating utility to aptitude. These placement decisions involve a rather small number of alternative treatments. An indefinitely large number of alternatives may be used, however, persons being located within smaller and smaller intervals on the scale. In the limit, the placement problem thus becomes a problem in measurement or estimation.

Even though an institutional decision maker thinks of his task as "prediction" or "measurement", he may use only a few discrete categories. An example is the fitting of teaching methods to the pupil's estimated IQ. One would not treat those with IQ 109 differently from those with IQ 110, but one could modify treatment with each appreciable difference. Such use of measurements is ordinarily better described as a problem in placement with adaptive treatment than as placement with fixed treatment, because past experience has determined

what treatment is most suitable at each IQ level and the decision maker is free to use that best treatment.

It is appropriate in this chapter to consider a second type of adaptation, namely the altering of quotas on the basis of test information. For instance, three rather different levels of training may have been developed, each involving its own course outline, test materials, and instructional aids; these programs would be difficult to change and so treatments would be fixed. But the proportion assigned to each treatment could be altered. Improvement in test validity would warrant assignment of more persons to the extreme categories. Adjustment of quotas may also be beneficial when treatments are adapted, if the maximum number of treatments is fixed. The discussion in Chapter IV regarding the optimal selection ratio deals with a rather similar problem.

RELATION OF UTILITY TO VALIDITY

Fixed-treatment conditions

To examine what utility results with fixed treatments and fixed quotas, we employ assumptions similar to those used with selection. Appendix 1 presents the detailed mathematical argument. There are $\underline{n}$ different treatments, fixed a priori, and we assume that for each treatment there is a linear function relating payoff to test score. The score continuum is divided into $\underline{n}$ intervals each bounded by y_t^I and y_t^{II} , which are fixed so that the proportion of the distribution falling in each segment equals the quota set for the corresponding treatment $\underline{t}$. With each treatment is associated a value of $\sigma_{e_t} r_{ye_t}$, the slope of the payoff function. The gain in utility per man tested is

$$\Delta U (\text{fixed}) = \sum_t \sigma_{e_t} r_{ye_t} \Delta \xi_t - C_y \quad [10] (1.19)$$

Here, $\Delta \xi_t$ stands for $\xi(y_t^I) - \xi(y_t^{II})$; below, we shall also write $\Delta \phi_t$ for $\phi(y_t^I) - \phi(y_t^{II})$. At some points it will be helpful to write B_{yd} for $\sum_t \sigma_{e_t} r_{ye_t} \Delta \xi_t$.

It is possible to rewrite [10] in terms of the differences between validity coefficients for adjacent treatments:

$$\Delta U = \sum_t \xi(y_t^I) [\sigma_{e_t} r_{ye_t} - \sigma_{e_{t-1}} r_{ye_{t-1}}] - C_y \quad [11]$$

The bracketed term is the covariance of score y with the difference in payoff under two adjacent treatments; these covariances are weighted by $\xi(y^I)$ to obtain the total benefit. Clearly, since the test may have a different validity coefficient for each treatment, utility is not a simple function of any single validity coefficient. However, the greater the correlations of the test with the differences

in payoff between treatments, the greater is the utility from using the test for placement.

It will simplify some of our discussion to express utility in terms of an underlying aptitude $\underline{s}$ as was done in connection with adaptive selection. Since we are restricting ourselves to a single test score $\underline{y}$, we can without further loss of generality introduce an $\underline{s}$ such that for every treatment $r_{ye_t} = r_{ys}r_{se_t}$. Since $m_{st} = \sigma_e r_{se_t}$,

$$\Delta U = r_{ys} \sum_t m_{st} \Delta \xi_t - C_y \quad [12]$$

In placement with fixed treatments and fixed quotas, utility is linearly related to the component r_{ys} of validity.

Adjusted quotas. Total utility is increased by adjusting the quotas even when treatments are fixed in advance. Strategy should ideally be adjusted so that each person receives that one of the treatments for which his expected payoff is greatest. The most profitable cutting scores are determined, for particular values of $\sigma_e r_{ye_t}$ and e_{ot} , from equation (1.22). The cutting scores are related to r_{ys} . As r_{ys} approaches zero, the optimal cutting scores depart from the mean; an increasing proportion of the cases are assigned to the particular fixed treatment which is best suited to average men. This treatment is also the one to which all persons should be assigned for the optimum a priori strategy; we designate this treatment as t_o . (This is not necessarily the t_o discussed in adaptive selection.)

$$\Delta U = r_{ys} \sum_t m_{st} \Delta \xi_t + \sum_t (e_{ot} - e_{ot_o}) \Delta \phi_t - C_y \quad [13] (1.23)$$

The second term here is less than zero. As in selection, the simple formula for gain in utility with fixed-treatment placement [10] includes a component which could be obtained merely by a priori adaptation, giving the treatment best for unselected men to everyone. For that reason, the gain from testing expressed in [13] is less than the gain with the same treatments and fixed quotas. The total utility including a priori utility is, however, greater with adjusted quotas than with fixed quotas.

In equation [13] the values of $\Delta \xi$ and $\Delta \phi$ depend on r_{ys} , and therefore ΔU with adaptive quotas is not a simple linear function of r_{ys} . The function depends on the particular treatments to be used.

Benefit compared to selection. Fixed-treatment selection is obviously a special case of fixed-treatment placement in which there are just two treatments, for one of which $m_t = 0$. Making this substitution reduces [10] to [2]. There may be combined selection-placement problems such that one group of

men is rejected while the others are assigned within the institution on the basis of the same test. These too can be treated as special cases of [10].

When just two treatments are in use, equation [11] becomes identical to [2], save that $\sigma_{e_t} r_{ye_t}$ is replaced by $\sigma_{(e_t - e_{t-1})} r_{y(e_t - e_{t-1})}$. Figure 5, presented in our discussion of selection (p. 34), therefore also applies to the relation of utility to validity in placement with two fixed treatments. It is also to be noted that the discussion of two-stage sequential selection in Chapter V and Appendix 4 applies to sequential placement into two categories, since the covariance of y with differential payoff may be substituted for $\sigma_{e_t} r_{ye_t}$ in all equations. It is important to emphasize that this paragraph refers to the gain in utility with testing.

Validation of tests for placement with fixed treatments

Utility depends on the differential payoff of the treatments as a function of aptitude. If all m_{st} were equal, the gain in utility as a result of placement would be zero regardless of r_{ys} . (See [12]). Placement using a univariate score is profitable only when the slope of the payoff function is different for different treatments. This does not necessarily mean that the test-criterion correlation r_{ye_t} must differ from treatment to treatment, since the slope m_{st} depends also on σ_{e_t} .

Although the implications of this fact for placement have not to our knowledge been discussed, a comparable result established in connection with differential prediction, i.e., classification, problems is well known (see, for example, (13, 71)). A test of general intellectual ability having equal validity for predicting success in engineering and in liberal arts does not predict in which field a student will do best. A test of mathematical ability which may be a poorer predictor for either curriculum assists in differential prediction because it is more closely related to engineering than to liberal arts. This relationship has been widely discussed in recent years in connection with both classification of workers and differential diagnosis of patients.

Differential payoff, rather than predictive validity as usually measured, is similarly important in placement. This implies that tests intended for placement purposes in schools may have been validated in the wrong way. Where two or more treatments depend on the same aptitude, testers have generally been satisfied to examine the validity of a test for predicting criterion differences within treatments instead of examining how well it predicts payoff differences between treatments, i.e., the interaction of outcome and treatment.

Use of a test in placement rests on the assumption that no one treatment yields greatest payoff over the entire range of scores in the population. Numbering the segments of the y continuum in order from 1 to n , and numbering the treatments correspondingly, we may rewrite [11] thus:

$$\Delta U = r_{ys} [(m_{st_2} - m_{st_1}) \xi(y_1'') + (m_{st_3} - m_{st_2}) \xi(y_2'') + \dots + (m_{st_n} - m_{st_{n-1}}) \xi(y_{n-1}'')] - C_y \quad [14]$$

The bracketed term summarizes the differential contribution of a perfect measure of s to placement, taking into account the particular treatments and the quotas for each treatment. This is multiplied by r_{ys} to determine the differential contribution of the test itself. Thus utility in placement depends on two things: the power of the test to measure the aptitude dimension s , and the power of s to predict differential payoff.

An aptitude dimension which is strongly related to criterion differences within treatments (all m_{st} being large) may have no relation to differences between treatments (if the m_{st} are equal). Consider a set of treatments for which two aptitudes are relevant, aptitude s_1 being estimated with validity 1.00 by test y_1 , and aptitude s_2 being estimated with validity .10 by test y_2 . This information alone does not permit us to say that test y_1 is superior. Even adding the fact that within all treatments s_1 is a much better predictor than s_2 does not settle the question. If $m_{s_1 t}$ is a constant for all treatments, y_1 has no value for placement. The importance of an aptitude for placement depends on the change in m_{st} over the treatments under consideration.

Publishers of scholastic aptitude or achievement tests frequently recommend those tests for sectioning students. The claim of empirical validity is supported by giving the correlation of the test with a criterion of success within some treatment, or occasionally within a pooled group from several treatments. It has not hitherto been recognized that what the consumer needs to know is the slopes of the payoff functions or other data on the interaction of test and treatment.

Tests presently used may be ineffective for placement even though they are good predictors within a treatment. Possibly quite different types of items would make superior placement tests, because qualities which determine differential response to various treatments are not generally those which best predict criterion performance within one treatment. General mental ability, for example, is likely to be correlated with success in mathematics no matter how the subject is taught. If the alternative teaching procedures are an abstract deductive method and an applied inductive method, the brighter students should do better with either approach. Payoff functions for both treatments against general ability will have positive slope, and may or may not intersect within the ability range. On the other hand, there may be other qualities of the individual (say, interest in abstract problems, or liking for rigorous reasoning) which would have quite different relations to the two treatments. A measure which predicted success under one treatment and not the other would be a much better aid to placement than a measure which predicts both.

Adaptive-treatment conditions

Where allowable, adapting the treatment to the persons assigned to it is expected to have advantages in placement as in selection. A priori, the best adaptation is to assign all men to the treatment best for the average man. A posteriori, the best treatment is determined from the general utility surface. Appendix 3 demonstrates that gain in utility is a function of the square of the validity with which the test measures the s dimension:

$$\Delta U = \frac{r^2 y_s}{4a} \sum_t \frac{\Delta^2 \xi_t}{\Delta \phi_t} - C_y \quad [15](3.6)$$

The parameter a reflects the curvature of the general utility surface for s.

Figure 24 shows the relation of ΔU to validity and quota, in an adaptive placement problem with two treatments. Here we assume that $a = 1$ and $C_y = .05$. Restricting the problem to two treatments means that there is a single cutting score y' between treatments, and

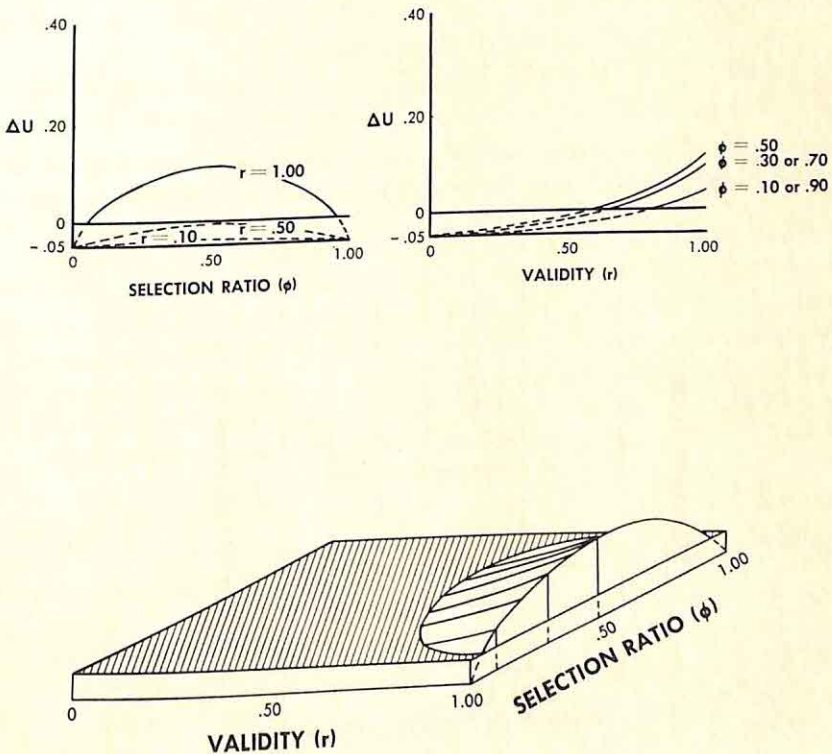


Figure 24. Relation of gain in utility to validity (r_{ys}) and selection ratio in adaptive-treatment placement

$$\Delta U = \frac{r_{ys}^2}{4a} \frac{\xi^2(y')}{[1 - \phi(y')][\phi(y')]} - C_y \quad [16]$$

Here, $\phi(y')$ is the proportion of men given treatment t_2 .

It is of interest to compare [16] with the corresponding equation [7] applying to adaptive-treatment selection. Equation [7] contains a term in r_{ys} . The reason is that "adaptive-treatment selection" is actually a mixture of fixed and adaptive placement: one group of men are given the treatment best suited to their average aptitude while the other group are given a fixed treatment (reject) for which $m_{st} = 0$. As a matter of fact, many mixtures of fixed and adaptive placement can be envisioned which are of practical importance. It is common practice to use a fixed treatment for the "normal" pupils in a school, for example, but to segregate handicapped pupils and give them a treatment adapted as well as possible to their capacities. Utility functions for such mixed situations will contain both r and r^2 terms.

Adjusted quotas. In Appendix 3 it is demonstrated that, with a specified number of treatments, cutting scores should be located so that for any pair of adjacent treatments

$$2y'_t = \frac{\Delta \xi_t}{\Delta \phi_t} + \frac{\Delta \xi_{t-1}}{\Delta \phi_{t-1}} \quad [17] (3.15)$$

The cutting score should be located halfway between the means of the groups assigned to adjacent treatments. If we plan to divide the group among n treat-

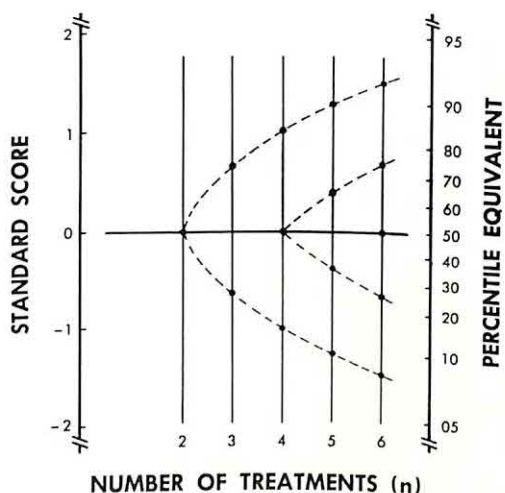


Figure 25. Optimum cutting scores for adaptive placement

ments, [17] provides $n-1$ simultaneous equations. Since y_t^1 is included in the argument of $\Delta \xi_t$ and $\Delta \phi_t$, the equations are not readily solved. It is of particular interest to note that in adaptive treatment the optimum cutting scores do not depend on r_{ys} . When there are just two treatments, the optimum procedure is to divide the group at the mean. With three treatments, the optimum cutting scores are +.65 and -.65. This allocates 26% of the cases to the two extreme groups, and 48% to the middle group. The optimum cutting scores can be determined for any value of n by successive approximations. These are indicated in Figure 25 for various numbers of treatments; the proportion of cases assigned to each treatment can be read from the percentile scale.

Desirable difference between treatments. When choosing treatments from a large number of possibilities, one can use similar or markedly different treatments for persons having a specified difference in aptitude. Differences of the form $m_{st_1} - m_{st_2}$ will be smaller in the former case than in the latter. For a fixed number of treatments, [17] defines the range of test scores within each treatment group. The expected average aptitude in each group can then be determined. The optimum value of m_{st} for the group is then determined from (2.5). This implies that there is an optimum separation of treatments for any given problem.

The greater r_{ys} , the more may adjacent treatments differ. With three adaptive treatments, 26% of the men (average test score = 1.20) should be assigned to the "upper" group. The treatment used should be that suited to men whose average aptitude is $1.20r_{ys}$. With low validity, all persons are assigned to treatments very similar to t_o . When y is a good measure of s , treatments may differ to a much greater degree.

When raters, assessors, or other judges estimate individual differences, they tend to overdifferentiate. For example, counselors tend to predict that more students will earn grades of B or better than actually do so. They ordinarily make insufficient allowance for error and therefore recommend differentiating treatment more than their predictors warrant. In the language of Chapter II, the assessor does not use the optimum strategy, which (for adaptive treatment) is indicated by [17]. This inefficient strategy is one factor contributing to the frequent finding that psychologists applying clinical judgment to objective data make poorer decisions than a statistical formula applied to the same data. Both empirical and mathematical aspects of such judgments have been discussed elsewhere (26).

Effect of increasing fineness of discrimination

We next inquire how utility varies when a test is used for a larger number of discriminations. When the number of different treatments n is increased, the decision maker faces a more difficult decision. It has generally been believed that a higher standard of validity is required for a test to be useful in making fine discriminations.

Fixed treatments, fixed quotas. An analysis of the change in validity for fixed-treatment placement can be made only under fairly specific assumptions. Suppose, for example, that our test is a measure of aptitude $\underline{s}$ and that we employ those $\underline{n}$ treatments from the general surface which would be optimal if r_{ys} were 1.00. Then, using (2.5) and substituting in [12],

$$\Delta U = \frac{r_{ys}}{2a} \sum_t \left[\frac{\Delta \xi_t}{\Delta \phi_t} \right] \Delta \xi_t - C_y \quad [18]$$

Now if we also assume that the same proportion of persons is assigned to each treatment, $\phi_t = 1/n$ and

$$\Delta U = \frac{r_{ys}}{2a} n \sum_t \Delta^2 \xi_t - C_y \quad [19]$$

If some other set of fixed treatments were used or some other quotas, the function would have a somewhat more complicated form. With uniform quotas,

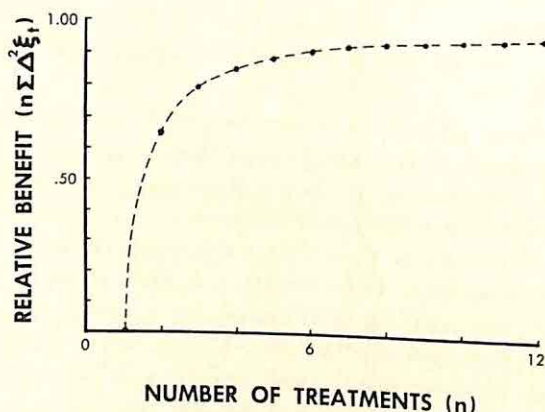


Figure 26. Increase in benefit from placement with increased number of treatments

the product $n \sum \Delta^2 \xi$ increases with $\underline{n}$ as shown in Figure 26. In the limit, as $n \rightarrow \infty$, $n \sum \Delta^2 \xi$ approaches 1.00.

Adaptive treatment. When equal numbers of persons are assigned to all treatments, and treatments are adapted to fit the group, [15] reduces to

$$\Delta U = \frac{r_{ys}^2}{4a} n \sum_t \Delta^2 \xi_t - C_y \quad [20]$$

Figure 26 also indicates the relation of $\underline{n}$ to benefit from adaptive placement with fixed equal quotas.

When both treatments and quotas are adapted, ΔU is related to $\underline{n}$ by a curve almost identical to Figure 26. With very fine division of the scale, i.e., as we approach continuous measures,

$$\lim_{n \rightarrow \infty} \Delta U = \frac{r_{ys}^2}{4a} - C_y \quad [21]$$

Under the conditions studied, the gain in utility as a result of increasing the number of treatments is small. After the first few treatments, benefit increases only very slightly with increased fineness of discrimination. With only two treatments, $n \Sigma \Delta^2 \xi$ equals .64, with four treatments, .86. Since the limit is 1.00 as $\underline{n}$ increases, the further increase is necessarily very gradual.

Everyone agrees that accuracy of measurement is advantageous, but three rather different types of advantage have been discussed in the literature. Keeping number of treatments fixed, we have already established that as measurement improves the resulting assignments achieve more nearly the maximum payoff from each individual. Secondly, we have seen that better measurement permits greater separation of treatments, with accompanying gains in utility. A third benefit sometimes expected is that better measurement warrants dividing persons among more categories; it remains to determine whether this is the case.

It has generally been believed that there is a strong relation between the accuracy of a test and the fineness of discrimination for which it may be used. The error of estimate based on r_{ys} specifies limits within which $\underline{s}$ will be expected to fall, and a more valid test estimates performance within any treatment more accurately. For this reason it has been suggested that greater accuracy permits one to place persons among a greater number of treatment categories. Bloom (7), for one, has interpreted test reliability in terms of placement decisions, requiring that frequency of erroneous placement shall not exceed 0.1%. By this reasoning, a reliability of .56 justifies use of three categories, .84 justifies five, and .96 justifies ten. The same approach could be applied to r_{ys} (which equals the square root of the reliability if $\underline{s}$ is equivalent to the "true score" on the test). The required coefficients would then be .75, .92, and .98 respectively, for three, five, and ten categories.

Our results do not support the conclusion that greater reliability or validity justifies division into more groups. Under the conditions leading to equations [19] and [20], finer differentiation results in the same proportional increase in benefit no matter what the value of r_{ys} . For example, the gain in utility with four treatments, disregarding C_y , is always 7.5% above that with three treatments. Our result differs from Bloom's because he evaluated decisions by the number of erroneous placements rather than by the magnitude of errors. We have counted gross errors of placement which lead to large changes in outcome as more serious than fine ones, whereas he has counted all errors as equal. Bloom's method of evaluating outcomes does not seem realistic for the

educational placement problems he discusses. The allowable number of subdivisions, we conclude, has nothing to do with the accuracy of the test.

The value of increasing the validity of the test as a measure of the $\underline{s}$ continuum is that it permits greater separation of treatments. With a good test people may be divided into ten groups and given ten sharply distinct treatments, i.e., treatments suited to quite different levels of aptitude. With a poor test people may still be divided into ten groups, but the ten treatments should not be extremely different. The conventional view is close to ours only when fixed treatments and adaptive quotas are involved. Suppose that there are six treatments, for example, those appropriate respectively where $\underline{s} < -2$, $-2 < \underline{s} < -1$, $-1 < \underline{s} < 0$, $0 < \underline{s} < 1$, $1 < \underline{s} < 2$, $2 < \underline{s}$. If quotas may be adapted, an increase in r_{ys} warrants placing more persons in the extreme categories. For r_{ys} below .40, the optimal strategy is to assign about 99.5% of the cases to the middle two categories, because that proportion of persons will have an estimated s_i between +1 and -1. Practically speaking, only two categories are used. As validity increases, more persons should be assigned to the second and fifth categories. When $r_{ys} = .60$ about 5% of the cases will go into each of these groups. Ultimately, with $r_{ys} = 1.00$, about 2% of the cases will go into the first category, 2% into the sixth, and about 15% into the second and the fifth. Thus the number of fixed treatments to which persons are assigned with appreciable frequency does increase with r_{ys} .

Under adaptive conditions, accuracy of measurement bears on the number of treatments used only as the decision maker may decide that small amounts of differentiation are not worth bothering with and combines two similar treatments into one.

THE INTERPRETATION OF VALIDITY COEFFICIENTS IN PLACEMENT AND MEASUREMENT

When Chapter IV examined the relation of test validity to utility of selection decisions, the evaluation of the validity coefficient in other uses of tests was held in abeyance. We may now summarize relations encountered in both selection and placement.

In placement, a test is used to locate the person among several treatments, and no one coefficient can be identified as the validity of the test. The test is likely to have a different correlation with criterion for each treatment. We have therefore distinguished three elements which determine payoff when prediction is made from a single test, namely σ_{et} , r_{ys} , and r_{set} . The variable $\underline{s}$ is common between a particular test and all treatments. So long as all tests under consideration are linked to the treatments through the same $\underline{s}$, we can develop a general function relating utility to the validity of the tests as a measure of $\underline{s}$, i.e., in terms of r_{ys} .

In this case, gain in utility in placement with fixed treatment and fixed quotas is a simple linear function of r_{ys} . This conclusion depends on the assumption that there is a linear relation of expected payoff to test score. Assuming in addition a second-degree surface relating $\underline{e}$, $\underline{s}$, and $\underline{t}$, the benefit in adaptive placement with fixed or adjusted quotas is a function of r_{ys}^2 . The results for fixed quotas apply to placement into any number of categories. Even if $\underline{n}$ becomes infinite, the linear or r^2 function applies, according to whether treatments are fixed or adaptive.

Typical decisions involving measurement are best interpreted as adaptive-placement problems. The person who "measures" in order to apply the treatment optimal for the obtained score is engaged in adaptive placement with an (hypothetically) infinite number of alternative treatments to be considered. Therefore the coefficient of determination r_{ys}^2 appears to be the proper index for evaluating the usefulness of a test in personnel decisions calling for numerical measurement.

The coefficient of forecasting efficiency has a long history, and has been widely used as an index of test merit. But in the light of these results we must now ask whether it has any proper place in test theory. If it is a proper index of test efficiency, there must be some decision problem and associated payoff function such that the coefficient is proportional to benefit from testing. The coefficient of forecasting efficiency evaluates a test by the absolute error of estimate, relative to the error of a chance estimate. This evaluation would be appropriate only in a problem where payoff declines linearly with error. One type of payoff surface (and to the best of our knowledge only one) can be described for which this is true, namely a ridge-shaped surface, formed by the intersection of two planes. In such a surface, the payoff function for any single treatment is like that in Figure 27, where the treatment shown is that optimal

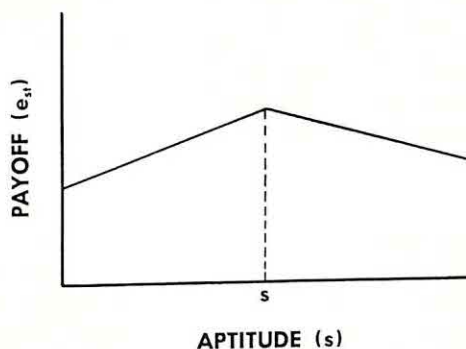


Figure 27. Possible payoff function for a single treatment

for aptitude $\underline{s}$. The coefficient of forecasting efficiency can be justified in a decision problem where the payoff function takes this shape, and the question therefore is whether such a function would occur in practice.

Outside the personnel field, measurement problems can be imagined where this function might apply. Suppose nails are sorted by size into bins, by a process involving some error. Then the carpenter who reaches for a certain size of nail for a given job will perhaps get one of incorrect size. The more the nail deviates in either direction, the worse suited it is: too small a nail will not hold the work firmly, too large a nail is ugly and risks splitting the wood. Perhaps the payoff function is neither linear nor sharply discontinuous, but over a range of sizes Figure 27 might be a reasonable approximation. Similar functions might be involved in the payoff from, say, various sizes of clothing allotted to a person, or various concentrations of electrolyte in a plating bath. Therefore we can argue that absolute error of estimate is a suitable loss function for some measurement problems.

But does it apply in personnel testing? At this point, we would say that it does not. Overestimating a score leads to assigning a person to a treatment he may be unable to profit from. But underestimating aptitude seems unlikely to have consequences like those of underestimating the size of a nail. The man who is placed in a treatment "too simple for him" will ordinarily outperform the man of lesser ability. The loss lies in the failure to assign the man to a treatment where he could do even better. In some situations boredom might cause inferior performance on the part of "excessively able" men, but we have not identified a situation where the payoff function seems likely to be sharply ridged; instead, the rising-then-declining function is likely to be a smooth, gradual curve. We therefore conclude that the coefficient of forecasting efficiency is not a suitable index for describing the value of a test in personnel decisions. We may have overlooked some type of decision which has a payoff function justifying this index, but it is surely not appropriate for typical selection and placement decisions. The person who proposes to use $1 - \sqrt{1 - r^2}$ as an index must bear the burden of proof, showing that this index does correspond linearly to gain in utility for some particular problem.

Recent writings have distinguished the value of a test for screening or selection from its value in other uses. The Brogden linear relation or the Taylor-Russell function were connected specifically to coarse screening, and have been regarded as inapplicable to precise decisions or prediction for an individual (see, for example, (70)). The coefficient of forecasting efficiency or the standard error of estimate is ordinarily invoked for these latter cases. The coefficient of determination r^2 is not often mentioned directly as an index of efficiency, although it is prominent in test theory.

Our results, taken all together, lead to the following quite different views:

1. There is no single "validity coefficient" on which the contribution of a test depends, save in the case of fixed-treatment selection. In a placement problem, a measurement problem, or an adaptive-selection problem, we can only compare the contributions of various tests measuring the same aptitude continuum. Statements can then be made about the value of the test as a function of r_{ys} .

2. In selection with treatment and quota fixed a priori, utility is a linear function of the correlation between test score and evaluated outcome, i.e., of the validity coefficient. In fixed-treatment, fixed-quota placement, utility is a linear function of the several validities corresponding to the various treatments. It is also a linear function of the validities of the test for predicting the differences in payoff between pairs of treatments. Finally, utility may be interpreted as a linear function of r_{ys} in fixed-treatment selection and placement.
3. When we may regard all successful men as making equal contribution to the institution, the Taylor-Russell tables are more appropriate for evaluating selective efficiency in fixed treatment than the linear function. Otherwise, the Taylor-Russell results are best regarded as a rough approximation to the linear relation.
4. In placement or measurement with adaptive treatment, utility is a function of r_{ys}^2 , which has the form of a coefficient of determination. To compare tests measuring different aptitudes, the parameter a of the payoff surface for each aptitude must also be taken into account.
5. Selection with adaptive treatment combines the features of fixed and adaptive placement. The relation of utility to validity involves both r_{ys} and r_{ys}^2 terms.
6. Where treatments are fixed, but quotas may be adapted according to the value of r_{ys} , the relation of utility to validity becomes too complex to be described in a simple index.
7. No case of practical importance has been found where the coefficient of forecasting efficiency is a suitable index of test efficiency.

7

EFFICIENT TESTING PROCEDURES

Among the chief problems of the decision maker is the efficient design of testing procedures. The time he can devote to testing is often severely limited, and how to use that time efficiently is a matter of concern.

ADJUSTMENT OF TEST LENGTH

One way in which the efficiency of testing can be increased is by adjusting a test or battery to that length which would yield maximum gain in utility for the decision problem in which it will be used. When a particular collection of items is used, gains in efficiency can be obtained by altering the number of items. With a battery of tests treated as a composite it may be desirable not only to alter the lengths of the separate tests, but also to omit some tests entirely despite the fact that they improve the multiple correlation.

Optimum length for a single test

The validity of a test rises as it is lengthened, the increase being gradual unless the units forming the test have low intercorrelations. Since validity does increase indefinitely, there might seem to be no limit to the desirable length of test. A utility analysis, however, demonstrates that there is an optimum length, beyond which increases in cost outweigh benefits from greater validity. This point was recognized by Hull.

Speaking of a battery composed by adding units with validity .40 and intercorrelation .20, he wrote:

The tenth [unit] adds only 1.2 points to the correlation yield. The question inevitably arises whether an increase of a single point or so in the correlation yield is worth the extra time and labor involved in giving and scoring an entire additional test unit. In any case it must be perfectly obvious that because of this law of diminishing returns a place must be reached sooner or later where the addition of a new test will not contribute enough to the prognostic value of the battery to justify the incidental expense involved. (43, p. 262)

Despite this early statement, the validity-cost balance has generally been ignored in discussions of test design, and has never been reduced to a definite function.

To study the effect of lengthening a test we express the general equation for placement or selection as follows:

$$\Delta U_k = B_{y_k d} - C_{y_k} \quad [22] \quad (5.3)$$

Here the subscript k indicates that we are considering a test of length k . B_{y_d} is the benefit from using test y in a particular decision problem d , and has been defined for each type of decision so far considered.

The cost of a unit test may be divided into two portions, C_0 and C_1 . The first element C_0 is an initial cost of testing which may be assumed to be constant, regardless of length. This "setup" cost takes into account assembling subjects, giving directions, etc. The marginal cost C_1 of a unit test may be assumed proportional to length, and includes examiner time, time of men tested, scoring costs, etc.

When k similar units are combined to form a test, benefit from testing is an increasing negatively accelerated function of k . Cost increases linearly and must eventually exceed benefit. The difference between benefit and cost is the net gain in utility.

The benefit must be specified for a particular unit test and a particular decision problem. As Appendix 5 shows, somewhat different equations ((5.4) and (5.13)) describe the increase in benefit with length in fixed-treatment decisions and in adaptive-placement decisions. Figure 28 shows the change in utility with length for both types of decisions. While these curves are based on a particular set of parameters ($B_{y_1 d} = 1$, $r_{y_1 y_1} = .30$, $C_0 = .05$, $C_1 = .02$), the general shapes of the curves would be similar for other parameters. Change

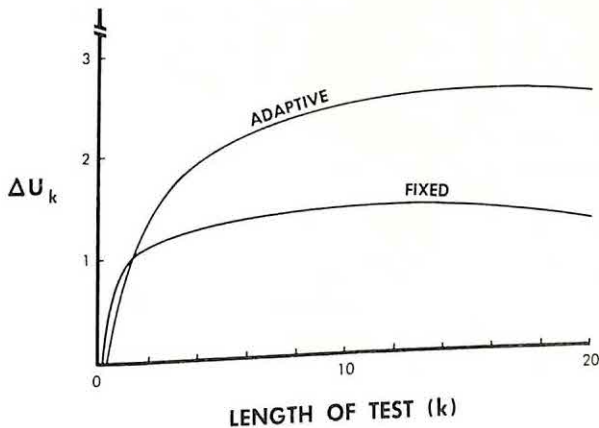


Figure 28. Utility as a function of test length

of parameters would shift the maxima and alter the curvature. No direct comparison of the curves for fixed and adaptive treatment is warranted, and no meaning is attached to the fact that the adaptive curve is higher in the figure, since the parameter B_{y_1d} is defined differently in the two cases.

For any test, there is some one best length (unless the test is too invalid to ever repay its cost). If the test is shorter than this optimum, the tester is not attaining full utility. As a test is lengthened beyond the optimum value, ΔU declines and eventually becomes negative. The utility curve is fairly flat over a large range of k , and it is therefore not critical to determine precisely the best k for a given situation. The left portion of Figure 29, based on (5.6),

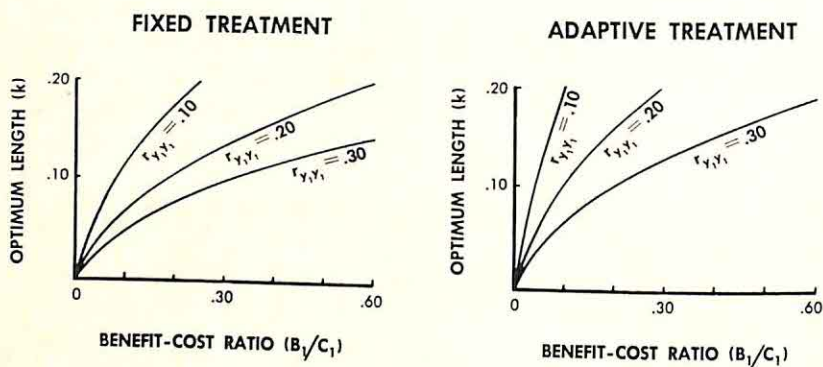


Figure 29. Optimum length of test as a function of other parameters

shows the optimum k for fixed treatment as a function of other parameters. The right portion (from (5.16)) shows the optimum for adaptive placement. In both cases, the optimum length of test increases as cost C_1 decreases, benefit B_{y_1d} increases, and/or intercorrelation $r_{y_1y_1}$ decreases. In selection, shorter tests should be used as the selection ratio departs from .50. The optimum length does not depend on C_0 , the "set-up cost". An optimum length of test for adaptive selection also exists (5.19). Adequate cost estimates are not available to indicate the range of B/C in typical practical situations. If the parameters employed in Figure 29 are realistic, many present tests are too long for greatest efficiency.

Optimum battery length

A similar effect occurs when tests are combined into a composite predictor, as is usual with test batteries. The multiple correlation rises slowly as the battery is augmented. In fixed treatment, the gain from adding the v th test is $\Delta R - C_{y_v}$, where ΔR is the resulting increase in the multiple correlation

and C_{y_v} is the cost of the y th test. A test which, used as a sole predictor, contributes sufficient validity to be worth its cost may not add enough in a battery to be profitable (cf. (5.10) and (5.11)).

Previous writers have suggested ways of determining optimum composition of a battery, when its length is fixed. Long and Burr (49) consider the case where one is dealing with tests whose lengths are fixed but not uniform. They show that the conventional procedure for choosing tests to form the battery should then be modified. They indicate how to select the best combination of tests for any specified testing time. Horst (39) goes further, altering the lengths of tests in order to maximize the multiple correlation for any fixed total testing time. His argument is parallel to ours save that he does not introduce a set-up cost for each new test. Taylor (63) provides a clear discussion of the Horst solution, with examples.

While these solutions consider the design of a battery with a specified length, they do not take the further step of determining the optimum length. As we have indicated above, the cost of additional testing at some point outweighs benefit from improved prediction. Using (5.10) and (5.11) in an iterative procedure similar to that of Horst, it is possible to build up a battery approaching the optimal as closely as desired both in length and composition for a fixed-treatment selection problem. Probably one could devise a more direct method, taking into account the type of decision problem, the cost of each test, its benefit at unit length, and all test intercorrelations.

The problem of maximizing the efficiency of a battery of tests can be further generalized to situations where many decisions are to be made from the obtained information. Such cases have been treated by Horst and others and will be discussed in the following chapter.

Distribution of effort in two-stage sequential testing

The tests used in a two-stage sequential plan should also be designed for maximum efficiency, the lengths of the first and second tests being adjusted according to their unit validities and costs. We deal with distribution of effort within a battery of fixed total length and cost; the optimal length of a two-stage battery has not been studied. No comparable problems arise in the usual multi-stage plan, where the tests used at successive stages are identical in validity and cost.

The ideal distribution of effort can be examined by computations similar to those in Chapter V. These computations are practical where the battery may be divided into units of equal validity and cost, but not for heterogeneous combinations. A given homogeneous battery may be divided in any proportion between the two stages of testing. For purposes of illustrating the effect of various divisions on validity, we have divided the battery into 20 identical units. Any number k of these units may be employed as the first screen. In this example, the intercorrelation of units is set at .10, C_1 at .005, and C_0 at 0.

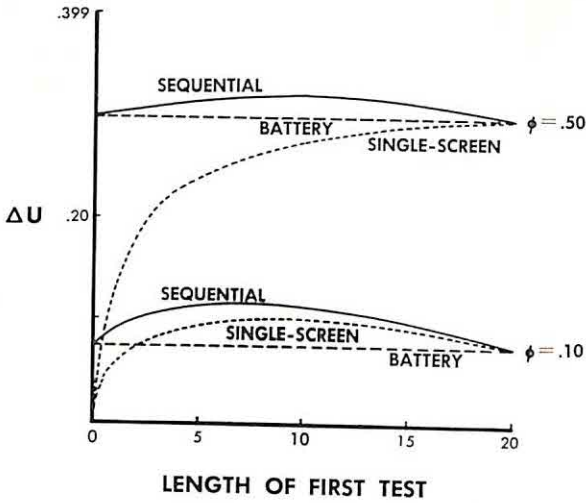


Figure 30. Utility in two-stage selection with varying distribution of effort

(As in Chapter V, utility and costs are relative to $\sigma_e r_{Ye}$.) Figure 30 shows how the gain in utility from Single-screen, Battery, and Sequential procedures is altered by changes in k , with selection ratios .50 and .10. The utility from Battery is constant, since the cost of the battery and the validity do not depend on the way it is divided. Utility from Single-screen increases with length in the manner of Figure 28. With the parameters here assumed, the first screen would profitably be lengthened beyond 20 units when the selection ratio is .50. The optimum length for the first stage of a two-stage battery, however, is about 10 units -- much less than the optimum length for the same test used as a single screen. For a selection ratio of .10 (or .90), the optimum single screen is one with k approximately 9, and the optimum length of first stage is about 8. If the first stage of testing is much shorter than this optimum, Battery is nearly as good as Sequential. If the first stage of Sequential procedure is much longer than the optimum, Sequential offers little advantage over the Single-screen. The proportion of the battery assigned to the first stage should decrease, other things being equal, when the selection ratio becomes more extreme, the intercorrelation of unit tests becomes higher, and cost per unit test becomes higher, or the product $\sigma_e r_{Ye}$ becomes less.

MULTIPLE-STAGE TESTING

A type of sequential testing more general than we have hitherto considered employs a large number of stages for individuals whose assignment is difficult

to determine, but for other individuals makes terminal decisions after one or a few stages. Such a multi-stage sequential testing process distributes effort very efficiently. The length of test is adjusted to the individual on the basis of the information about him as it is received.

Wald and the Columbia Statistical Research Group (referred to below as SRG) have developed multi-stage sequential strategies for a great variety of problems, including some of those which concern us. To date, multi-stage procedures have been given detailed attention only under the condition that the same test is used at every stage, i.e., when the decision maker continues at every stage to collect further samples of the same sort of information. Wald (69, p. 114 ff.) has developed a "recursion" formula which specifies the ideal sequential procedure and permits determination of the benefit from such a procedure, provided one knows the a priori distribution of aptitudes and the cost of testing. This method is laborious to treat computationally, and the older SRG methods may be regarded as an adequate approximation. The SRG methods are designed for situations where one lacks prior knowledge of the distribution of true scores, and where outcomes are specified in terms of risks of various sorts of errors rather than in terms of utility units. We have been able, however, to adapt the methods to certain of our fixed-treatment problems. For those readers interested in the procedures, we summarize the SRG methods in Appendix 6 and indicate how the optimal multi-stage strategy can be determined for any specified parameters. The present discussion confines itself to presenting typical results.

Where persons are to be divided into two groups, each to receive a fixed treatment, a desired division point y'_{∞} is established. Then, on the basis of the information y available after any stage of testing, one can estimate the probability that an individual is above y'_{∞} . When this probability approaches one or zero, one can confidently make a terminal decision regarding the individual. The optimal strategy provides two scores for each stage of testing which constitute the boundaries of the region in which testing continues. As soon as an individual crosses either of these boundaries, testing for him ceases. A person attaining the higher of these scores is assigned to one treatment, and a person falling at or below the lower score is assigned to the other treatment.

The utility from a multi-stage test depends on the cost of the unit test, the intercorrelation of unit tests, and the relevant payoff functions. To study the gain in utility from a multi-stage procedure, benefits and amounts of the gain in utility from a multi-stage procedure, benefits and amounts of the gain in utility from a multi-stage procedure were computed, using an ILLIAC program for multi-stage testing required were computed, using an ILLIAC program for multi-stage testing devised for us by Kern Dickman. The program applies to decisions with two categories. Here we discuss the results in terms of selection rather than placement.

As a matter of convenience, we have defined the unit test as one for which $r_{y_1 y_1} = .10$. In practice, one might employ larger or smaller units, with

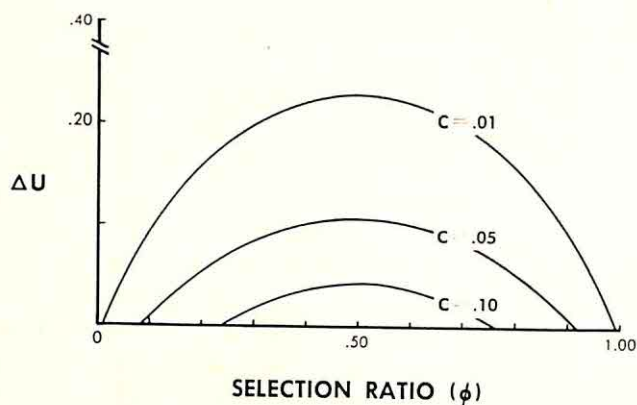


Figure 31. Utility in multi-stage selection as a function of selection ratio

correspondingly larger or smaller intercorrelations, but this change would affect results very little. Cost and utility are expressed relative to $\sigma_e r_{y_{\infty} e}$. At each true score total cost is proportional to the number of stages of testing, and net utility is determined by subtracting cost from benefit. For any cost of unit test and selection ratio, there is an optimum strategy. When persons are selected by this strategy, the utility varies with selection ratio in the manner shown in Figure 31. These curves are rather similar to those found for two-stage sequential testing. The relation of utility to cost of the unit test is depicted in Figure 32.

The advantage of sequential testing over non-sequential testing is demonstrated in Figure 33. Here, one curve shows the benefit from sequential testing assuming $C_1 = .01$. The other curve indicates the utility to be expected

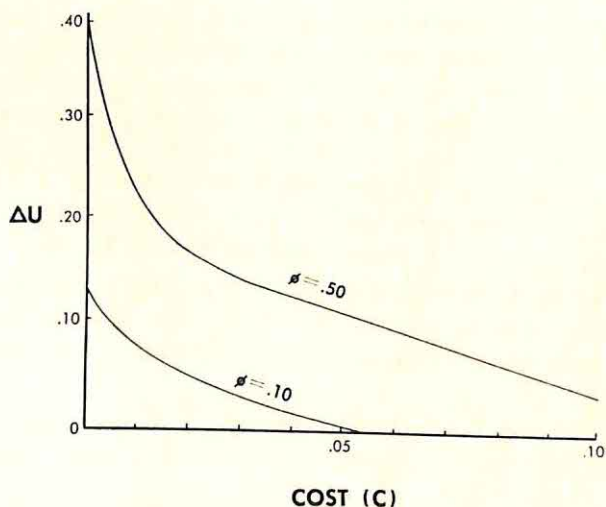


Figure 32. Utility in multi-stage selection as a function of cost

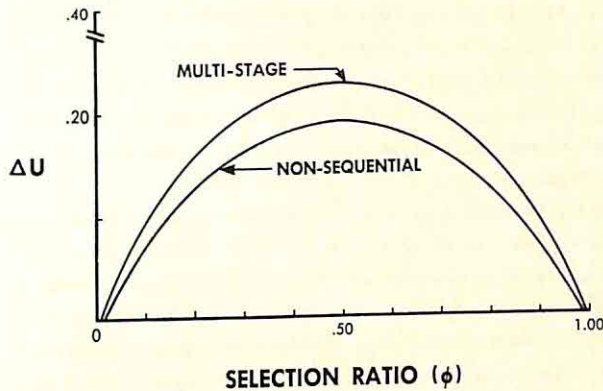


Figure 33. Utility from comparable non-sequential and multi-stage procedures

if every man were given a test of uniform length, the length of test used at any selection ratio being exactly equal to the average length of test for the sequential procedure. Calculations for the non-sequential curve are only approximate. The sequential plan where the same total amount of testing is distributed unequally over the men tested is superior at any selection ratio to the conventional procedure. More than this, because the lengths of the non-sequential tests are very near the optimum in each case, it is not possible by any adjustment of length to increase the utility of the non-sequential plan to the point where it equals the sequential procedure. That is to say, with this test the utility reached by sequential selection is unattainable by a non-sequential method.

Sequential testing for placement purposes

Sequential methods of obtaining information can be applied to placement decisions other than the two-treatment case discussed above. Sobel and Wald (59) describe a procedure for deciding which of three intervals along a continuum a parameter falls into. They employ assumptions such as those discussed in Appendix 6. This method can be applied directly to placement decisions in personnel work. While more general solutions involving greater numbers of categories and other assumptions can probably be developed, none has yet been described in a practically usable form. The principles of sequential placement are adequately illustrated by the Sobel-Wald treatment of the three-category problem.

No study has been made of the contribution of sequential placement for more than three treatments where payoff under each treatment is a linear function of score. The strategy specifies that testing will continue only if the total score after any stage of testing falls between certain values. A terminal decision is made for each person as soon as the risk of error is reduced to tolerable size. Testing is most extensive for a person whose true score falls near one of the boundaries. It is evident that considerable time can be saved in testing persons far from any borderline, but the saving will be small for

persons near the borderline. As the number of treatments increases, there are more borderlines and more persons must be given long tests. Sequential testing for placement into many categories, therefore, appears likely to have limited value as the number of categories becomes large. Indeed as $n \rightarrow \infty$, i.e., in a measurement problem, it has been established that the sequential method has no advantage (62, 72).

Designing tests according to a sequential pattern can perhaps improve their efficiency for other than selection or placement problems. While most test theory assumes the test to be identical for everyone, some individual tests have involved decisions at one stage as to what questions will be administered next. Thus the Stanford-Binet Scale employs the vocabulary score to determine a trial basal age. In measurement of sensory thresholds, the ascending and descending trials of the Method of Limits are similarly sequential in nature. Recently developed answer sheets which permit immediate scoring of responses open up the possibility of sequential procedure in group tests.

One situation where such a procedure promises to be valuable is in analytic proficiency or achievement measurement. Here, many different types of attainment need to be checked so the subject can be given further training where necessary. If a minimum performance standard can be established for each type of attainment, a short test can be administered for each. When these tests are scored, decisions can be made at an acceptable level of confidence regarding those dimensions where the person's score is extremely high or low, and he can be asked to take a further test on each dimension where his score is less extreme. This procedure is repeated until a decision has been made regarding every dimension. The total amount of testing required for any individual is considerably less than would be needed if sufficient items were administered non-sequentially to make sure that every decision reaches the same minimum level of confidence. Somewhat more complicated variants of the procedure can be used for multidimensional tests where the tester is interested in identifying the person's highest aptitude or his salient personality characteristics, etc.

Attention may be drawn to a paper by Somerville (60) which appears to represent a first step toward the mathematical study of sequential processes where each stage of decision making indicates what hypotheses should be tested next. Somerville studies a two-stage sampling problem where the first stage obtains data on many dimensions, and only the dimension with the highest estimated mean is to be tested in the second stage. In this problem, he is able to determine the optimum length for the first stage.

There is another possible use of sequential technique in designing single-score tests. In two-category placement, it has sometimes been found desirable to use a "peaked" test where the difficulty of all items is chosen so as to be maximally discriminative at the critical aptitude level. With more categories, the efficiency of testing could perhaps be improved by adjusting the difficulty of items according to the person's performance on the preceding items. For

example, if there are three categories, separated by aptitude s' and s'' , there could be two levels of item difficulty. Level 1 would be such as to discriminate most sharply between persons above and below s' ; level 2 would discriminate at s'' . The test might begin with five items at level 1, which would be scored immediately. Persons who pass (say) three or more would be directed to try a group of items at level 2, while the remainder would take further items at level 1. At the end of this stage, those successful at level 1 (all items considered) would move to level 2. Those who had attempted level 2 would be divided, some moving back to level 1, others remaining at level 2. Those who, in the end, succeed at level 2 are assigned to the top category, and those who fail at level 1 to the bottom category; the remainder go into the middle category. Such a procedure can be varied by introducing more categories and more levels of difficulty, by altering item intercorrelation, by altering the number of items at each stage, and so on.

We have made only a sketchy exploration of this procedure. It appears that such a sequential procedure is advantageous only when the items of the test are highly homogeneous in content, i.e., have high tetrachoric intercorrelations. Then, if the items are sufficiently homogeneous in difficulty within levels, and sufficiently widely spaced between levels, the various subtests will discriminate between individuals on the basis of their position on the aptitude continuum. In general, the higher the tetrachoric correlations, the greater the number of levels which may be used and the closer together may be the division points. It is well known (50) that item intercorrelations in educational and psychological tests are rarely high enough to yield "difficulty factors", unless they are applied to a group having an exceptional range on the attribute measured. Very high intercorrelations are presently encountered only in a few unusual instruments, such as Guttman-type attitude scales. These devices would seemingly be benefited by further development of sequential test designs.

A sequential procedure for administering a scholastic aptitude test was described by Krathwohl and Huyser to the 1956 meeting of the American Psychological Association. Preliminary results (unpublished) indicate that the time required for obtaining estimates of ability can be substantially reduced by this method.

8

THE BANDWIDTH-FIDELITY DILEMMA

In deciding whether to use a test, the practical worker considers not only its validity but also its range of applicability. There is an obvious difference between the value of a measure of mathematical proficiency, used in several decisions about a student by his teachers and counselors, and that of an equally valid measure of drawing ability applied to only one or two decisions. The way a test is scored affects its value; a single overall score will usually have a narrower range of application than a pattern of sub-scores on the same test. Utility analysis suggests that the contribution of a test be judged over all decisions, rather than in terms of validity for any one decision. It is this total contribution to the institution which determines which tests, or which scores, should be used.

An example will show the significance of this argument. Instruments available to classroom teachers for evaluating pupil adjustment, such as questionnaires, sentence completion tests, and sociograms, have limited but positive validity. A pupil's poor score on such an instrument can influence many of the teacher's decisions: decisions regarding how to discipline him, how much pressure for greater achievement to apply, what social activities to suggest, and whether to make a diagnostic case study of him. Although the adjustment inventory improves any single decision much less than a test of, say, spatial ability improves the decision to which it is relevant, the former test, applicable to many important decisions, is likely to be more beneficial.

Working with a person over a period of time, a teacher or therapist applies a great variety of treatments. Each time, he classifies the person as requiring (ready for) the treatment or not. We may regard the test interpreter as having many hypotheses of the form: "Treatment t_A is suitable for this individual". A decision to accept or reject is made regarding each hypothesis. These are typically placement decisions, since the person not given the treatment remains within the institution. These decisions are independent whenever one decision does not affect the others. We shall term a set of such independent decisions, a compound decision. Our consideration of the contribution of a test over decisions will be limited to this case.

A model involving independent decisions is admittedly oversimplified, since decisions about a person are likely to be integrated rather than independent. The decision of a counselor to recommend remedial reading for a student may be independent of his decision to try to broaden the student's participation in social affairs, but it is more likely that recommending remedial reading will dictate postponing expansion of the student's social life. This example might be regarded as allowing four alternative treatments (modify reading and social program, modify reading alone, modify social program alone, modify neither). If the probabilities that a person will be given the four treatments are predictable from the probabilities that he will be given reading and social treatments, considered separately, we have a compound placement decision (i.e., two independent placement decisions). If decisions are not independent, the problem involves four categories. While these categories might all be predicted from a single test, it is much more likely that the appropriate predictors will involve at least two factors, and therefore this problem is no longer one of placement. Such general classification problems are more common in industry and military personnel work than are compound placement decisions. Even where decisions are not strictly independent, however, the compound placement model may be a useful approximation.

Our inquiry involves two sub-questions. First, it is necessary to find an expression for the value of a test or battery used for compound decisions. Secondly, we employ this expression to identify general principles regarding the optimum design of the testing battery in problems involving compound decisions.

UTILITY OF A TEST IN COMPOUND DECISIONS

We recall that the gain from making any one decision $\underline{d}$ is

$$\Delta U_d = B_{yd} - C_y \quad [22]$$

B_{yd} is the benefit obtained when test y is used as a basis for this decision, and depends on the character of the decision problem, the treatments, and the location of cutoffs, as well as on the validity of the test for whatever aptitudes are involved.

If the same test, scored in the same or a different manner, is used for several decisions about the same persons, further benefit is obtained without further cost. The grand total contribution of testing is

$$\sum_d \Delta U_d = \sum B_{yd} - C_y \quad [23]$$

The contribution of the test is more or less proportional to the number of

decisions for which the set of scores can be used. A test which applies to v similar fixed-treatment decisions contributes the same net utility as another test having v times as much validity which applies to only one. In adaptive treatment, where B_{yd} is a function of r_{ys}^2 , a test must apply to v^2 similar decisions before it is as profitable as another test having v times as much validity for one such decision.

One must be cautious in implying that tests can compensate for poor validity merely by being used repeatedly. When the validity of a test is near zero, multiplication is most unlikely to raise the total utility to a satisfactory level. Moreover, a positive validity coefficient computed on one sample does not guarantee usefulness for a test. Sampling error of low correlations is substantial, and our argument is based on coefficients established for the population.

DISTRIBUTION OF EFFORT IN A MULTI-SCORE BATTERY

The test designer and the user of tests frequently have to choose between careful estimation of a single variable and more cursory exploration of many separate variables. Tests may be constructed to yield separate scores on a number of diverse, internally homogeneous scales, or to provide a single measure loaded with the general factor underlying items. Particularly where the test is to be used in a variety of decisions rather than to predict one single criterion, questions arise as to whether to establish independent part scores or to obtain a careful measure of a single attribute (48). The person choosing published tests for a testing program faces similar questions, since he can use available time to measure one or two variables by means of long tests, or employ a much larger number of short tests measuring a variety of characteristics.

This dilemma may be described in the language of the communications engineer as a choice between "wideband" and "narrowband" tests. In using a particular channel, such as a telegraph wire, one may either crowd many messages into a period of time, or give a single message slowly and repetitively. The former, more varied message has greater "bandwidth". The wideband signal transmits more information, but the clarity or dependability of the information received is less than for the narrowband signal except under ideal communication conditions. Random errors can seriously confuse the wideband signal; this is spoken of as a lack of fidelity. The tester's situation is analogous. If he concentrates on facts relevant to a single decision, he gets a much more dependable answer than if he spreads his effort. But by concentrating, he leaves all his other questions to be answered on the basis of chance alone.

This suggests that in any decision situation there is some ideal compromise between variety of information (bandwidth) and thoroughness of testing to obtain more certain information (fidelity). For the purposes of designing

electronic communication circuits, such ideal compromises have been worked out within the Shannon mathematical theory of communication (57). Because of its close analogy to testing problems, we once expected the Shannon theory to provide a basis for test design. Upon close examination, the Shannon model does not fit the tester's problem (23, 24). Though information theory is suggestive, the tester's problem must be treated within the more comprehensive mathematical structure of decision theory.

Previous solutions. We wish a maximally efficient strategy for gathering information when a large number of decisions are to be made. These decisions depend on various aptitudes, which in the general case may or may not be correlated. Some relatively general mathematical solutions have been offered, but none of them leads to easily comprehended or communicated results. Certain restricted solutions have appeared in the literature. Elfving (32) and Chernoff (18) suggest a way to determine the optimal distribution of effort which minimizes the total squared error of estimate. Chernoff shows that this least-square index leads to the "locally optimum" division of effort under a wide range of conditions, provided error of estimate is small. The Elfving-Chernoff method assumes that all decisions are equally important, and does not make allowance for a set-up cost C_0 . Horst (42) has employed similar assumptions to obtain an optimal battery for fixed total testing time, where the battery is to predict various criteria.

These investigators have assumed that the tests are intended to estimate the standard score on each criterion as accurately as possible, there being a different criterion for each decision. The index used is the sum of the squared errors of estimate for the various criteria, or what is equivalent, the sum of R_{yc}^2 over criteria, where R is the multiple correlation of the test scores with the criterion. Such an index clearly applies to measurement problems where treatments will be adapted to the estimated criterion scores of the individual (cf. eq. [21]).

Placing its entire emphasis on the validity coefficients as it does, this index is not a fully satisfactory basis for judging the contribution of a test. When all criterion measures are reduced to standard scores, they are in effect assumed to be equally important. Secondly, the index leaves quotas or cutting scores out of account. While these are irrelevant in measurement with an infinite number of categories, the quotas become an important factor in utility when tests are used to select among a few alternative treatments. Finally, it is to be noted that the least-squares criterion is relevant to adaptive-treatment decisions, but according to our previous chapters is not a suitable index for fixed-treatment decisions. This index would, however, agree with our analysis in warning against judging test merit by the validity coefficient against each single criterion considered separately.

Our treatment will take into account the parameters of the decision problem. In so doing, however, we find it necessary to impose other serious restrictions so that we deal with only certain special cases.

Equivalent decisions and tests. The first special case to be considered is that where each of a set of terminal decisions depends on a different aptitude, and there is a test which measures each aptitude. We assume that $B_{y_1 d}, r_{y_1 y_1}, C_0$, and C_1 are the same for every test, i.e., that the decisions and the unit test relevant to each are similar. The correlations among tests and among outcomes are zero. Both set-up and marginal costs are assumed to depend entirely on time, but set-up costs are independent of the length of the test. Bandwidth is described by $\underline{v}$, the number of dimensions to be tested.

The question is: If the tester has $\underline{T}$ units of testing time available, is he wiser to divide that time over all decisions or to concentrate on measuring one variable? Or is his best choice intermediate between these? When he lengthens one test, he improves its contribution but is left with many decisions to be made on a chance basis; if he uses many short tests he can cover all his decisions, but with highly fallible information.

With equivalent tests and similar decisions, the tester should divide his time equally among whatever number of tests he gives. This is true because the relation of benefit to length (eq. (5.1)) is convex, and therefore the contribution corresponding to length $\frac{1}{2}(k_1 + k_2)$ is always greater than the average of the contributions at k_1 and k_2 . The problem reduces to determining the optimum number of tests of uniform length. (If decisions and tests were not uniform, it would be desirable to use tests of different lengths and no simple statement about bandwidth could be made.)

In fixed treatment, the contribution of any test as a function of length is as was indicated in Figure 28. The constraint on testing time fixes a total allowable cost C_T . Appendix 7 demonstrates that the contribution from a battery of $\underline{v}$ uniform tests, each of length $\underline{k}$ and each used for a separate fixed-treatment selection or placement decision, is

$$\sum_d \Delta U_{kd} = v B_{y_1 d} \sqrt{\frac{r_{y_k y_k}}{r_{y_1 y_1}}} - C_T \quad [24] \quad (7.2)$$

where

$$v = \frac{C_T}{C_0 + k C_1} \quad [25] \quad (7.9)$$

If C_0 is zero, $v = C_T / k C_1$. Substituting this value of $\underline{v}$ in [24], it can be seen that utility is a monotonic decreasing function of $\underline{k}$. Utility then increases indefinitely as $\underline{v}$ increases (although $\underline{v}$ could never increase beyond the point where each test contains only one item). If there were no initial cost it would be profitable to increase bandwidth up to the limit of the number of decisions to be made.

Where there is an initial cost, there is a limit to how short the tests can be because extremely short tests cannot repay the set-up cost. The relation of gain in utility, for all tests combined, to the number of tests has the general form shown in Figure 34. (As a matter of convenience, ΔU and $\underline{v}$ are both expressed relative to C_T in this figure. C_0 , C_1 , and B_{y_1d} are set at .05, .02, and .2 respectively.)

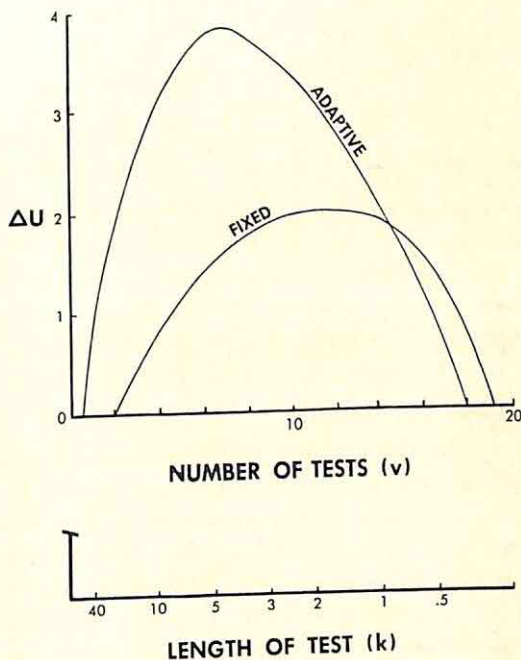


Figure 34. Relation of utility to bandwidth

If the uniform decisions involve adaptive-treatment placement, the equation analogous to [24] is

$$\sum_d \Delta U_{kd} = v B_{y_1d} \frac{r_{y_k y_k}}{r_{y_1 y_1}} - C_T \quad [26] (7.14)$$

This function also is shown in Figure 34.

The optimum length of any one test (which, substituted in [25], gives the optimum bandwidth) is obtained by maximizing [24] or [26] with respect to $\underline{k}$. For fixed treatment under the specified conditions the optimum length is

$$k = \frac{1 - r_{y_1 y_1}}{4 r_{y_1 y_1}} \left[-1 + \sqrt{1 + 8 \frac{C_0 r_{y_1 y_1}}{C_1 (1 - r_{y_1 y_1})}} \right] \quad [27] (7.13)$$

For adaptive-treatment placement, the optimum is located where

$$k = \sqrt{\frac{C_0 (1 - r_{y_1 y_1})}{C_1 r_{y_1 y_1}}} \quad [28] (7.25)$$

It is to be noted that with the postulated uniform decisions the optimum bandwidth does not depend upon $B_{y_1 d}$, but only on the cost and intercorrelations of unit tests. Even more important, the optimum number of tests has no relation to the number of decisions to be made (save as this provides an upper limit). If the optimum $\underline{v}$ is five in a given situation the decision maker should measure five dimensions, whether he is concerned with five decisions or one hundred. Moreover, we can arbitrarily define the length of a unit test, say by adjusting it to make $r_{y_1 y_1} = .10$. The optimum therefore depends entirely on the ratio C_0/C_1 for the tests under consideration.

Such exploration as this demonstrates that it is frequently profitable in making independent decisions to divide time among several tests rather than to devote all time to a single test. Too great a dispersion of effort, however, is just as unprofitable as too great a concentration. With uniform tests, greater bandwidth is profitable for independent, uniform decisions when there is low initial cost C_0 relative to C_1 . This condition will obtain when the introduction of new dimensions can be accomplished without allowing time for new directions to the subject, practice items, etc. Personality questionnaires and interviews meet this requirement and therefore can increase bandwidth profitably; in ability measurement, set-up time for new dimensions is sufficient to place greater restriction on bandwidth. The foregoing analysis assumes, however, that we cannot single out certain dimensions as being particularly important, an assumption which is contrary to fact in most selection and guidance.

Non-equivalent, independent decisions. The values of $B_{y_1 d}$ vary from test to test if the unit-test validities differ, or if the decisions to be made from the tests are not equally important, or involve different cutting scores, different numbers of treatments, etc. It may be seen from (5.6) and (5.16) that the optimal k_d increases as $B_{y_1 d}$ increases, other parameters being equal, for either fixed or adaptive treatment. In general, the greater the disparity in values of $B_{y_1 d}$, the more time should be given to the tests with larger $B_{y_1 d}$ even though this reduces bandwidth. This principle requires further qualification, however, when $r_{y_1 y_1}$ or the ratio C_0/C_1 differs from test to test.

Appendix 7 indicates a general solution for the optimum distribution of effort in adaptive placement which takes into account all parameters of the unit test provided one defines the unit test by fixing C_1 . Then, designating the length of any one test as k_d , the gain in utility over a set of v tests is

$$\sum_{d=1}^v \Delta U_d = \sum_d B_{y_1 d} \left[\frac{r_{y_k y_k d}}{r_{y_1 y_1 d}} \right] - \sum_d (C_{0d} + k_d C_1) \quad [29] (7.14)$$

The total number of test units $K = \sum k_d$ is defined by

$$K = \frac{C_T - \sum C_{0d}}{C_1} \quad [30]$$

A composite parameter $D_{y_1 d}$ is introduced.

$$D_{y_1 d} = \frac{B_{y_1 d} (1 - r_{y_1 y_1 d})}{2C_1 r_{y_1 y_1 d}} \quad [31] (7.17)$$

Then, for any specified set of v tests, the optimum length of the test for the decision d' is given by

$$k_{d'} = 1 - \frac{1}{r_{y_1 y_1 d'}} + \frac{D_{y_1 d'}}{\sum D_{y_1 d}} \left[K - v + \sum \frac{1}{r_{y_1 y_1 d}} \right] \quad [32] (7.21)$$

There is a similar equation for each other test. Appendix 7 discusses an iterative procedure for comparing the utility achieved using one set of tests, with their optimal lengths, to that from another set of tests. No comparable explicit formula for k_d can be obtained for fixed treatment, because the system of equations (7.6) of the appendix has not been solved.

Horst's procedure for finding the optimal group of subtests and their appropriate length is also an iterative method, but allows the tests and criteria to be correlated so that a change in length for any one test may affect the utility for all decisions. With this added complication, the optimum lengths cannot be expressed by an explicit formula such as [32] and must be determined by successive approximations.

Our special case gives a somewhat clearer picture of the factors affecting distribution of effort. Increase in $B_{y_1 d}$ (relative to C_1) for one particular decision increases k_d . Increase in $r_{y_1 y_1 d}$ for any d markedly decreases k_d . When tests are alike with respect to these parameters it is best to divide time evenly among them. As one test takes on greater validity or importance, other things

remaining equal, it is better to give more time to that test. As the benefits from the unit tests become more unequal, a point is reached where it is more profitable to spend all available time on one test. For a dimension where unit tests have low intercorrelations, benefit increases greatly with length; but where the intercorrelations are high, there is little value in lengthening the test.

Koopman (47) has made a general study of distribution of effort between two experiments. He introduces an exponential "return function" relating benefit to length of test, which is comparable to our function based on the Spearman-Brown formula. Despite this difference, his conclusions are consistent with ours.

We present a hypothetical numerical example in order to illustrate our conclusions. Suppose that there are five tests, with these respective values of B_{y_1d}/C_1 and $r_{y_1y_{1d}}$ at unit length: 5, .30; 1, .30; 0.5, .30; 0.5, .10; 0.1, .10. We assume C_0 to be negligible and let $K = 20$. Computing from [32] gives these respective lengths: 11.2, 3.7, 2.0, 5.6, and -2.5. The negative value implies that test 5 yields insufficient benefit to be used when $K = 20$, so it must be dropped from the set. When the optimum lengths are recomputed, with $v = 4$, the values become 10.3, 3.3, 1.7, and 4.6. The values of the parameters in this example were deliberately made markedly unequal. At unit length, test 1 gives ten times as much benefit as test 3, and five times as much as test 2. In this situation their optimum lengths have the ratios 6:1 and 3:1. Test 4 and test 3 contribute the same benefit at unit length but because the unit-test intercorrelations of test 4 are smaller, it has a much greater optimum length. The lengths of the tests are ordered according to the ratio of B/C_1 to $\bar{r}$.

Generalizations. In establishing generalizations, we pool the information derived from our analysis, and also from the work of Horst and Koopman. These results apply to a single-stage battery used as a basis for terminal decisions.

1. It is generally profitable to divide testing time among many tests rather than to concentrate on a single test, when many decisions are to be made regarding the same persons.
2. Using fewer tests than there are decisions may be preferable to dividing time over all decisions. Such restriction on bandwidth becomes important under any of the following conditions: (a) the initial costs of the tests are unequal and not all negligible, (b) the contributions of the tests at equal length are markedly unequal, or (c) intercorrelations of unit tests (units being defined by equal cost C_1) are higher for some of the tests than for others.
3. For any given problem there is an optimal distribution of effort, both with respect to number of tests to be given and amount of time to be devoted to each test. Where the tests are uniform, it is wise to divide time equally among the tests given. Where the tests are

not uniform, it is profitable to allow time to the tests roughly in the order of the ratio of their contribution per unit cost to the intercorrelation among unit tests.

4. It is not critically important to employ precisely the optimum length of test. Minor departures from the optimum reduce utility very little in the cases we have examined.

Our analysis confirms that there is indeed an optimal compromise between "bandwidth" and "fidelity" in testing for compound placement decisions. Similar results are to be expected for classification decisions. The validity coefficient of a test relative to a particular decision describes only one aspect of its usefulness. Where it is in competition with other tests, the amount of time to be given to it depends upon its contribution per unit cost and the correlation among units. The cumulative benefit from a series of moderately valid tests may outweigh the benefit to be expected from a smaller number of more dependable tests.

CLASSIFICATION DECISIONS

Classification decisions, unlike those previously discussed, are necessarily based on multivariate information. The problems associated with classification are considerably more difficult than those encountered in selection and placement, and our treatment is consequently much more limited.

At present, one of the most crippling difficulties is that the ideal strategy for a classification decision involving quotas cannot be determined. A strategy has two aspects, the information-gathering procedure and the procedure for allocating men after the information is obtained. A truly optimal strategy can be designed only by considering both aspects simultaneously, and this presents mathematical difficulties of a high order, as Dwyer (29) has pointed out. The sequential methods of Magwire which we shall discuss below offer an ideal strategy for classification except when there are numerical quotas to be satisfied exactly in a given sample. Where sample quotas are fixed, it is presently possible to describe an ideal allocation strategy only after the test battery has been specified.

While the classification problem for fixed treatments can be described in algebraic terms, the formula describing utility contains an expression which cannot be integrated and is therefore not simply comprehended. Numerical substitution may be used to explore the functions relating utility to characteristics of the test battery, but so many parameters are involved in even the simplest classification decisions that little clarification is gained thereby.

We have therefore limited our discussion to an interpretation of previous work on classification in the language of decision theory. Our discussion brings together a variety of hitherto isolated results, and also raises many important questions. We first describe the fixed-treatment classification problem in a form comparable to that for selection and placement, in order to examine what factors affect utility. The adaptive-treatment problem is discussed in a less technical fashion, our aim being to examine the meaning of adaptive treatment in a classification problem. The proposals of Horst (41, 42) regarding the design of classification or differential-prediction batteries are

then considered in terms of the assumptions about the decision problem which they appear to invoke. Finally, we consider possible development of sequential methods in classification.

CLASSIFICATION WITH FIXED TREATMENTS

The most common classification problem is one in which certain fixed treatments are available, and each man is to be assigned to one of them. Each man i yields a certain payoff e_{it} under whatever treatment t he receives; with n treatments, there are n values of e_{it} for him. Some writers formulate an assignment strategy in terms of operations upon the n by n matrix of estimated e_{it} . We, however, shall consider the test scores rather than the estimated payoffs as the starting point, to obtain a formulation readily applicable to large samples.

A set of tests $y_a, y_b, \dots$ is available. We assume that payoff is a first-degree function of scores on these tests. For each treatment t , there is a linear combination of scores Y_t determinable by multiple-regression methods which best predicts the payoff under that treatment. The adequacy of this prediction is indicated by the multiple correlation $R_{Y_t e_t}$. The several Y_t need not be uncorrelated. (For the reject treatment, payoff is assumed to be zero regardless of the test scores, but this function is merely a degenerate case of the ordinary regression equation.)

A vector notation ($\tilde{y}_i$ or $\tilde{Y}_i$) is used to indicate the particular pattern of scores belonging to person i ; for each $\tilde{y}$ there is a corresponding $\tilde{Y}$. When treatment t is applied to a man having a given pattern of predictor scores, the estimated payoff is

$$e_{\tilde{Y}_i t} = \sigma_e R_{Y_t e_t} Y_{it} + e_{ot} - \sum_y C_y \quad [33]$$

Any strategy for assignment states a complete set of values of $p_{t/\tilde{Y}}$ or $P_{t/\tilde{Y}}$. Regardless of whether the strategy employed is optimal or not, we can express as follows the total utility which is attained by using it:

$$U = \sum_t \sigma_e R_{Y_t e_t} \int_{\tilde{Y}} p_{\tilde{Y} t/\tilde{Y}} Y_t d\tilde{Y} + \sum_t P_t e_{ot} - \sum_y C_y \quad [34]$$

The integration is over all values of $\tilde{Y}$. In the second term, P_t is the proportion of persons assigned to the treatment. This equation applies to either fixed or adaptive quotas. In an optimal strategy, all persons having the same $\tilde{Y}$ are assigned to the same treatment; $p_{t/\tilde{Y}}$ is 1.00 for one treatment and zero for all others.

For each treatment there is a payoff function having the form of [33]. These equations describe hyperplanes which intersect pairwise in hyperplanes of one

less dimension. The intersections divide the space defined by the $\tilde{Y}$ dimensions into regions such that for each region there is one particular treatment which gives the highest obtainable payoff for every person in the region. Where there is no quota constraint, these intersections therefore indicate the optimum strategy.

Where quotas must be satisfied, the hyperplanes must be shifted parallel to themselves in such a way that the adjusted regions contain the desired number of persons. Cardinet (17) has developed a procedure for graphic determination of this strategy where there are just three treatments. From the estimated payoffs for the three treatments, the first centroid factor is extracted. The scores for each person on the two remaining orthogonal components are plotted. Three lines in this space indicate the locus of points where two of the estimated payoffs are equal. These would serve as boundaries of regions in the quota-free case. If fixed numerical quotas are to be satisfied, one counts the number of persons in each region. After observing which regions have too many persons and which have too few, one shifts the boundary lines in the appropriate direction parallel to themselves; to facilitate this step, Cardinet draws the lines on a plastic overlay which slides to the new position. A count is made of the number of persons added to or subtracted from each region by each shift. The position of the lines which satisfies the quotas is quickly located. This visual method will not serve where the number of treatments and the number of information dimensions are both greater than three.

Cardinet's method is particularly helpful when quotas are stated in terms of the proportion of a population to be assigned to each treatment. If the procedure is once applied to a large sample from the population, the lines which cut off the desired proportions specify the strategy to be used on all further samples, so long as the nature of the population remains constant.

The strategy assigns all persons in a certain region to t_1 , and the integral in equation [34] is equivalent to the integral of $p_{\tilde{Y}} Y_t d\tilde{Y}$ over that region. Y_t will not in general be normally distributed within this subgroup even when Y_t is normally distributed over all individuals. The reader may confirm this by considering the very simple case of three treatments, the third being reject and the composite scores Y_{t_1} and Y_{t_2} being uncorrelated. Suppose further that $\sigma_{e_{t_1}} R_{Y_{t_1} e_{t_1}} = \sigma_{e_{t_2}} R_{Y_{t_2} e_{t_2}}$ and that the quota $p_1 = p_2$. Then to treatment t_1 will be assigned those persons at the upper end of the Y_{t_1} distribution, excepting those for whom $Y_{t_2} > Y_{t_1}$. Because of this last requirement, the distribution of Y_{t_1} for those assigned to t_1 will not be normal. The lack of normality of the Y_t distribution makes it impossible to simplify the equations for utility in classification.

To study the gain in utility and the factors which determine it, we now restrict ourselves to the fixed-quota case. The a priori utility is $\Sigma p_{t_{ot}} e_{ot}$ and the gain in utility from testing is

$$\Delta U = \sum_t \sigma_{e_t} R_{Y_t e_t} \int_{\tilde{Y}} P_{\tilde{Y}} P_{t/\tilde{Y}} Y_t d\tilde{Y} - \sum_y C_y \quad [35]$$

Writing $\bar{Y}_t$ for the average value of Y_t among those assigned to treatment t ,

$$\Delta U = \sum_t \sigma_{e_t} R_{Y_t e_t} p_t \bar{Y}_t - \sum_y C_y \quad [36]$$

Equation [10], which described utility in placement with fixed treatment and fixed quotas, is a special case of [36], since $\Delta \xi_t$ in [10] equals $p_t \bar{Y}_t$. Whereas [10] could be evaluated directly from the tabled normal distribution, it is difficult to evaluate [36]. For three treatments it is possible to make use of the Pearson tables of the bivariate distribution, but the computations for even this limited case are laborious. We have not evaluated equation [36] for various values of the parameters, but we may review some of the findings of Brogden (13) who computed, under certain special assumptions, the average utility per man selected.

Brogden considers the selection-classification problem where some men are to be rejected and the remainder are divided among several treatments. He determines the value of decisions based on a single "general" predictor (equally correlated with outcome under every treatment) and compares this to the value obtained by using a separate "differential" predictor for each criterion. The general predictor permits rejection of the poorest men, but the men it accepts can be allocated to treatments only by chance. The differentially-scored battery predicts the outcomes for the person under the several treatments and therefore, insofar as quotas permit, one can assign him to the treatment which promises the largest outcome from him. Brogden assumes zero intercorrelations among the predictor scores Y_t obtained from the differential battery, and assumes that the correlation $R_{Y_t e_t}$ between an outcome and its corresponding predictor score is the same for all treatments. Other assumptions include first-degree payoff functions, equal σ_{e_t} , and a normal distribution of Y 's. The quotas are equal for all treatments other than reject.

The value of each type of battery is determined for varying numbers of treatments. Figure 35 shows the benefits obtained when the number of treatments for accepted men varies from 1 to 5, the number of predictors in the differential battery increasing accordingly. The quota for each treatment other than reject is fixed at 20%. The validity for the general predictor with respect to each treatment is taken as .50, and each validity coefficient $R_{Y_t e_t}$ of the differential battery is .50. Where just one treatment is available for accepted men, the differential predictor is identical to the univariate one. With two treatments in addition to reject, 40% of the group is accepted. The average quality of men accepted is necessarily lower than in the first case, but the decline in ΔU is much less serious with the differentially scored battery than with the single predictor. As the number of treatments is further

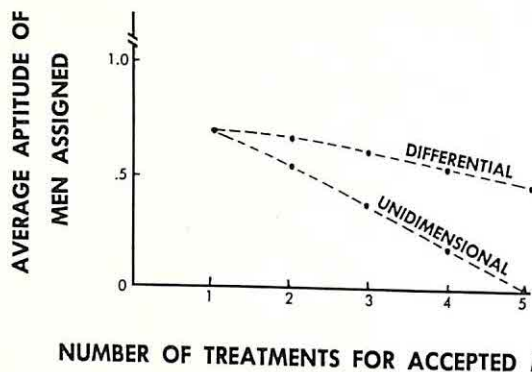


Figure 35. Quality of men assigned with unidimensional and differential information

increased the advantage of the differential battery becomes even more apparent. Brogden's example, one should note, assumes that differential predictors have the same level of validity as the general predictor, which is not true for differential batteries so far developed. These mathematical results, however, encourage efforts to improve differential predictors, since their potential advantage is great.

In Brogden's results it is notable that gain in utility has no simple relation to the predictive validity of the tests. The univariate predictor has the same validity as any of the differential predictors, yet the differential battery yields the greater gain in utility. Equation [34] indicates that, as we have previously found for other fixed-treatment fixed-quota decisions, utility is a linear function of the validity coefficients. This relation, however, holds only when the assignment strategy is fixed independently of the test validities. If the validity of the battery with respect to any one treatment is altered, the payoff surfaces intersect at new positions. This should lead to a change in the assignment strategy and thus change the value of $\bar{Y}_t$. The net change in utility is therefore a quite indirect function of the change in validity.

To clarify the relation of utility to differential validity, we can express equation [36] in terms of orthogonal variables. The tests $y_a, y_b, \dots$ may be resolved into orthogonal components $y_j = y_1, y_2, \dots$. While any set of orthogonal components can be used for mathematical purposes, it is perhaps worthwhile to think of extracting first those components which account for the r_{ye_t} . Any components of the test battery which are uncorrelated with outcome will then disappear from the estimation equations. Equation [36] holds when the Y_{it} are predicted from the estimated y_{ij} instead of from the original test scores.

Under the assumption of a first-degree payoff surface,

$$\tilde{y}_{it} = \sigma_{e_t} \sum_j r_{ye_t} y_{ij} + e_{ot} - \sum_y C_y \quad [37]$$

We may introduce a special notation for the purpose of considering differential validity. Let t' be the treatment to which the man is assigned and $\bar{y}_j(t')$ be the average y_j among those assigned to t' . Then, by a development analogous to that for [36], we find that if each man is assigned to the treatment optimal for him

$$\Delta U = \sum_{t'} \sigma_{e_{t'}} \sum_j r_{y_j e_{t'}} p_{t'} \bar{y}_j(t') - \sum_y C_y \quad [38]$$

For any y_j , $\int p(y) y_j = 0$ and therefore over all t and j , $\sum_t p_t \sum_{t'} \sigma_{e_{t'}} \sum_j r_{y_j e_{t'}} p_{t'} \bar{y}_j(t') = 0$. This expresses the gain in utility which results from assigning persons at random to all treatments in the proportions p_t , and is of course equal to zero, according to the definition of a priori utility. Subtracting this expression from the right member of [38] and rearranging, we obtain

$$\Delta U = \sum_{t'} \sum_j p_{t'} \bar{y}_j(t') \sum_t p_t (\sigma_{e_{t'}} r_{y_j e_{t'}} - \sigma_{e_t} r_{y_j e_t}) - \sum_y C_y \quad [39]$$

This equation is analogous to [11] for gain in utility in placement. The term in parentheses is the difference in the slopes of the payoff functions for the specified pair of treatments with respect to a particular y_j . This equation demonstrates again the well-known fact (71) that the value of a measure for classification depends not on its ability to predict outcome within treatments, but on its interaction with treatment. We remind the reader, however, that the hyperplanes bounding the region where persons are assigned to t' shift as the validity shifts; therefore, even when the quota remains constant, $\bar{y}_j(t')$ varies with the validity. This complicates the relation between utility and validity.

The discussion of implications on pages 67-68 applies to classification as well as to placement. As was mentioned there, the importance of differential validity has been discussed previously, but the role of σ_e has been too little recognized. For a test (or a factor) to be useful in dividing men between two treatments, it is necessary that $\sigma_e r_{y_e}$ for treatment t_1 differ from the corresponding product for t_2 ; inequality of the correlations does not suffice. For example, suppose one predictor dimension has validities .20 and .60 for two treatments. This might seem to promise differential value, but if the respective σ_e are 3 and 1, the predictor is not useful for making assignments. Conversely, we might have thought that a factor has no differential value if its correlation with outcome is .40 for each treatment. If the variance of the two payoffs is different, however, making assignments on the basis of this factor may yield appreciable gains in utility. Analyses of outcomes expressed in standard-score form have led to oversimplified discussions of the value of a classification test solely in terms of validity coefficients. In a given practical situation, a difference of one standard deviation in outcome will not have the same value for every treatment. The importance of considering the slope of the payoff function in utility units is therefore obvious.

It will be noted that the estimation of payoffs in this section has been based on regression formulas, and that the intersection of regression surfaces has been used to determine the boundaries of assignment regions. We have avoided use of the well-known Fisherian discriminant functions because the underlying model of the discriminant function appears unsuitable for personnel decisions. The basic concept underlying the discriminant function is that a number of classes exist, to each of which a particular individual "belongs". Differences among members of a given class are disregarded. This function divides the score-space into regions so as to maximize the probability of "correct" classification of each individual. It is possible to weight errors of a given type or to specify acceptable Neyman-Pearson risks for each of the possible types of misclassification. The regression model, on the other hand, is consistent with the view that there is no uniform cost of assigning a person "belonging in category A" to category B. Instead, the cost depends upon the pattern of scores. The theory of the geneticist is based on discrete classes such as species, and it is meaningless for him to speak of different payoffs from allocating different members of the same class in the same way. Here, the discriminant function is fully appropriate. The model is much less appropriate in personnel decisions where there is no theory of qualitatively different types of persons. Even in clinical diagnosis where categorical labels are commonly used, the discriminant function is probably unsuited. Calling a person "schizophrenic" means that a certain treatment will be applied. One expects variation in the response of these "schizophrenics" to the treatment, a variation which is in principle predictable by means of a payoff function. This variation should be taken into account in developing the classification strategy.

The foregoing section may be summarized as follows. By assuming that payoff under any treatment is a first-degree function of test scores, we arrive at an equation for gain in utility which shows that the value of a testing procedure depends in a complex manner on the quotas, the several validity coefficients for each test, and the dispersion of payoffs under each treatment. This relationship is a general one, of which the expressions relating gain in utility to validity in selection and placement are special cases. The general case is far too complicated to justify such a generalization as that increases in validity bring a proportionate increase in benefit from testing. Laborious calculations for specific sets of parameters would be required to establish the true shape of such functions in classification.

CLASSIFICATION WITH ADAPTIVE TREATMENT

In both selection and placement, the concept of adaptive treatment led to significant findings. We therefore have inquired how adaptive-treatment classification could be defined, even though we do not go on to develop formulas for this case.

It has been customary in job classification, in clinical diagnosis, etc., to think of each treatment as discrete and qualitatively different from the alternatives. While qualitative differences distinguish such pairs of treatments as lobotomy and psychotherapy, in many other personnel decisions we may think of a continuous multidimensional manifold of treatments (cf. pp. 24-25). Thus, while parole and continued imprisonment seem at first glance to be qualitatively different, one may conceive of a continuum between freedom and strict incarceration. Intermediate stages include relaxation of supervision in prison, furloughs, and parole with mandatory reporting. The full range of possible treatments can be characterized only by considering many dimensions; besides degree of liberty, there can be variation in the amount of education and rehabilitation offered, the degree of sympathy displayed by supervisors, and other aspects of the treatment.

This concept of adaptation is relevant to most areas of applied psychology where treatments have hitherto been viewed as discrete and unordered. Teaching methods certainly may vary by degrees along such continua as amount of active practice, rapidity of pacing, and amount of explanation. Between jobs there is often a similar series of intermediate possibilities combining the features of the several jobs in various degrees. In the Strong Vocational Interest Blank, for example, Chemist and Author-journalist seem like divergent vocational paths, but there exist intermediate opportunities such as Science reporter or Editor of chemical reports. Strong's factor analyses of interests indeed may be regarded as identifying some of the continua along which occupations (i.e., treatments) differ.

Whether treatments can be adapted will depend on the situation. Often, although it is impossible to adapt treatment to each individual, it is possible to adapt to the average level of aptitude in each group. Therefore, as in selection (see p. 44), the ultimate aim of research becomes the discovery of the best combination of treatment categories and classification procedures, within whatever practical constraints exist.

In order to examine adaptive classification formally, it would be necessary to introduce aptitude dimensions $s_1, s_2, \dots$. This could be done simply by defining each $\underline{s}$ as the "true score" on one of the orthogonal $\underline{y}$ dimensions previously discussed, so that $r_{y_e t} = r_{y s} r_{s e_t}$ for each corresponding $\underline{y}$ and $\underline{s}$. A person's pattern of $\underline{s}$ scores is independent of the lengths of the tests used to measure his $\underline{y}$ pattern. In placement, the treatment was characterized by the slope of the payoff function with respect to a single $\underline{s}$ dimension. When there are several $\underline{s}$ dimensions, the payoff function for each treatment has a slope with respect to each of these dimensions. The nature of the continuous payoff surface involving all treatments and all dimensions is therefore much more complicated than in placement. Since at present we have no empirical facts about even the simple surfaces involved in adaptive placement, it would be unduly speculative to consider what form the surface for a classification problem

might take. Ultimately it should be possible to choose reasonable assumptions and to derive the relations between utility, the characteristics of the decision problem, and the tests employed.

DESIGN OF TESTING PROCEDURES

Single-stage batteries

How to design a test battery which will be maximally efficient for the classification of personnel is a problem of great importance, for which only Horst has offered solutions. By examining what assumptions about the decision problem his model entails, we can provide a partial basis for evaluating them. His two proposals deal with constructing the best batteries for "multiple absolute prediction" (42) and "differential prediction" (41). We have already seen (p. 91) that the former procedure is designed to select that subset of tests which will maximize $\sum_t R_{Y_t}^2 e_t$. This is an appropriate index of gain in utility in compound adaptive placement, where several treatments will be given to each individual, each treatment being adjusted to his ability on one dimension. Horst's other proposal is to select tests which maximize the sum of the squared "differential validities" of the test battery. A differential validity coefficient indicates how well the given tests can predict differences in payoff between two treatments, i.e., it is the correlation between the difference in payoff and the best "least-square" estimate of that difference obtained from a test battery. Horst shows that for standardized criterion scores, this index of differential prediction is equivalent to maximizing the difference between the average predicted variance and the average predicted covariance of criterion scores. By simple analogy to the compound adaptive placement problem, we can immediately place one interpretation on the Horst "differential prediction" procedure. This procedure minimizes the sum of squared errors of estimates of criterion differences, whereas the "absolute prediction" procedure minimizes the same errors for the original criteria. This formula then does arrive at the best battery for a series of compound decisions, where each decision depends on the measured difference between two criterion scores. Every pair of differences is taken into account successively in determining the series of treatments for the individual.

This appears not very similar to the usual classification problem. Only in quota-free adaptive placement (measurement) was the gain related to the squared error of measurement of a single variable; Horst's index therefore appears most likely to apply to adaptive classification where every person can be treated differently. Certainly his procedure does not provide the ideal battery for fixed-treatment classification. In this case, as [36] makes clear, utility is a function of the R_{Y_e} rather than their squares, and the boundaries between regions must enter into the determination of utility. The Horst solution, moreover, makes no adequate provision for a reject group who receive

no courses, job assignments, etc. Thus his analysis would apply only when all individuals tested are to be utilized.

Perhaps a generalized payoff function for an adaptive-treatment problem could be found such that gain in utility would be proportional to the sum of squared errors which constitute Horst's index. This would require that the payoffs under various criteria have equal variances, that the criterion dimensions have equal weight, and that the summed cross-products of errors be unimportant. It appears that only in a very special case will Horst's index be exactly proportional to gain in utility.

Horst has offered the only systematic procedure for maximizing the efficiency of a classification battery, and thus has taken an important pioneering step. The function used to define efficiency does not correspond clearly to any common type of decision problem, and it is demonstrably not the correct function for the fixed-treatment example to which Horst applies the method. Further work may show that Horst's method is a useful approximation for cases where some other efficiency index would be preferable on logical grounds. Until such evidence is available, we cannot have confidence that the battery developed by Horst's method will be maximally efficient for a particular decision problem.

Sequential testing

Sequential methods should be especially advantageous in classification because a large number of dimensions are usually involved and exhaustive measurement is out of the question. Often a limited amount of information will serve to eliminate a number of treatments from consideration, and to indicate one or two of the dimensions as especially critical in choosing among the remaining treatments. If, for example, a general aptitude battery consisting of a series of short tests shows that person i stands high on the aptitudes required for jobs A and B, the next stage of testing can concentrate on whatever aptitudes differentiate A from B. For another person having a different pattern on the general battery, quite different tests would be relevant to the final decision. Such sequential testing in which the second stage is tailored for the individual should be more efficient than a non-sequential plan.

The basic theory for multidimensional sequential testing can be derived from an unpublished paper by Magwire (52). He was investigating "the sequential choice of experiments", but his methods appear to be directly applicable to testing of individuals. He states the decision problem in a general form. A number of tests are available. After any stage of testing, the individual may be assigned to a fixed treatment or may be held for further testing. At each stage of testing the person's expected payoff under each treatment can be estimated. If the payoff under one treatment is greater than would probably be obtained by testing him further, thus arriving at a more certainly correct decision but incurring additional costs of testing, a terminal decision is made. Wherever it is profitable to test further, Magwire's procedure indicates which

test should be given next. The procedures developed are a generalization of Wald's recursion formula (69), and take into account the payoff functions and the cost of testing.

From Magwire's work it appears possible to develop an ideal sequential strategy for the quota-free fixed-treatment case which is often encountered in guidance or diagnosis. A strategy can also be developed for the case where population quotas are fixed, provided that the population distribution is known. Where a finite numerical quota must be exactly satisfied in a particular sample, however, Magwire's solution would not apply. At this time, no use of Magwire's formula in actual computations has been reported, and the computations may prove far too laborious for practical use. Even so, the theory opens the way to generalizations regarding the design of sequential procedures and may permit the development of computational shortcuts and approximations.

10

EVALUATION OF OUTCOMES

The assignment of values to outcomes is the Achilles heel of decision theory. Once outcomes have been evaluated, one can proceed in a fully rigorous fashion to compare particular decisions or general strategies. The evaluation of outcomes, however, seems often to be arbitrary and subjective, leading one to question whether any of the conclusions from decision theory can be trustworthy if the starting point itself is open to dispute.

The most telling answer to this objection is to point out that decision theory invokes no more subjective evaluation than does any method of arriving at courses of action. Every choice between actions involves evaluations, and every doctrine or set of principles embodies value judgments. Decision theory is no more dependent on evaluation than is traditional measurement theory or discriminant analysis. The unique feature of decision theory or utility theory is that it specifies evaluations by means of a payoff matrix or by conversion of the criterion to utility units. The values are thus plainly revealed and open to criticism. This is an asset rather than a defect of this system, as compared with systems where value judgments are imbedded and often pass unrecognized.

It is always difficult to set down a payoff matrix assigning comparable values to all the consequences of a decision, as those who have applied game theory to military strategy have discovered. Comparing the value of bombing an enemy city to the value of preventing a flood at home seems absurd because, at first glance, these events are incommensurable. The comparison is required, however, by the fact that a single act (e.g., shifting engineers from dam construction to airfield construction) enhances the probability of one outcome at the expense of the other. Whoever decides on such an act is weighing the "incommensurables" on the same balance, whether he does so consciously or not. Personnel decisions likewise require a balancing of seemingly noncomparable outcomes. The personnel manager may let a humanitarian outcome such as the self-respect of an aging worker offset tangible losses in production. A school may select students on the basis of factors (e.g., religion, social class) which have no relation to probable academic achievement; in so doing,

the policy maker is allowing outcomes other than achievement to compensate him for the fact that he is not securing the best possible students.

In clinical decisions, the outcomes include the duration of treatment, the amount of staff effort absorbed, and the contribution (or cost) of the patient to the community after release. It is at this point that the "balance sheet" concept seems discordant with the humanitarian purpose of the institution. The decisions made, however, must consider the community welfare. "We had better take this patient in, to relieve strain on his family"; "We had better use our space on cases where there is greater prospect of recovery"; "We cannot invest this much therapeutic time on a single patient"; "This treatment does not cure, but it makes the patient more manageable" -- all these statements reflect an intention to obtain maximum advantage for minimum cost. Nor is this incompatible with concern for the patient's welfare. If it wishes, the institution may include the patient's expressed feeling of well-being in its evaluations. The critical point is that the institution must still balance this against other outcomes; cheerfulness purchased at the price of delaying the patient's becoming independent may be a bad bargain.

Much attention has been given to the logic of evaluation and to procedures for making estimates of value (see 1, 4, 58, 65), but these difficult problems have not been handled as successfully as the choice of strategy once values are assigned. The progress to date consists largely in defining the problem of evaluation and distinguishing among different approaches to the problem. We shall review these possibilities as they might be applied to personnel decisions.

Attention was drawn in Chapter II to differences between "institutional" and "individual" decisions. In the former, a single decision maker is concerned with a large number of decisions of the same sort, and a reasonable meaning can be given to the concept of the total utility of a set of decisions, since all the decisions are evaluated by the same payoff matrix. The individual decision, on the other hand, must be evaluated by the individual's personal payoff matrix, and the same person rarely confronts the same decision repeatedly. There have been attempts to formulate strategies which take into account simultaneously the welfare of many individual decision makers. Such "welfare economics" cannot be developed save by the introduction of some superevaluation which provides a rule for balancing one man's satisfaction against another's dissatisfaction (4, 65). Any one such rule can be defended against another only in terms of the judgment of some superordinate decision maker. It follows that the individual decision can be considered only as an isolated event, for which any concept of "averaging outcomes" over many decisions is inappropriate. Since additive utilities have been assumed throughout our previous chapters, it is most important here to consider evaluation for institutional decisions.

INSTITUTIONAL EVALUATION OF OUTCOMES

Comprehensiveness of evaluations

Although it is a truism to say that in evaluating a decision all outcomes must be taken into account, the principle is deserving of some discussion. Employment decisions are commonly validated solely against some measure of the proficiency or rate of performance of the men accepted. There are, however, many other consequences: some number of beneficial suggestions or acts of leadership, some costs of training and supervision (see 27), some degree of absenteeism and some spoilage of materials.

The variety of consequences of a decision is well illustrated by the studies of Raines, Hunt, and associates (44, 54) who attempted to determine the value to the Navy of a neuropsychiatric screening program. Among the consequences for which actual criterion data could be obtained were frequency of subsequent psychiatric discharges, frequency of bad conduct discharges, amount of time spent in hospital, and frequency of disciplinary offenses. In addition, they point out that there are probable differences between screened and unscreened men in proficiency and efficiency of job performance.

In estimating the cost to the military service of utilizing marginally adjusted men, the assumption is sometimes made that this cost involves solely those men who are discharged before they complete their required period of service. It is taken for granted that, if a marginal recruit manages to complete his enlistment and to receive an honorable discharge at the end of his term, his service is *ipso facto* successful and he has demonstrated his worth to the service. Unfortunately, completion of service without discharge is no guarantee of the quality of the service rendered.

... These results . . . show that such [marginal] men are more 'expensive' to the services, and that when their utilization is demanded by manpower needs it will be necessary to make provision for the added demands they will entail upon medical and disciplinary facilities. They also confirm the clinical picture of the maladjusted individual as one who, even when he is meeting the formal adjustment standards of his group, is doing so at a greater cost to the group's medical facilities and with greater friction upon its social organization than is his adjusted compatriot. (54, pp. 12-13)

It is important to note how this detailed analysis differs from the evaluation employed by Taylor and Russell, or Berkson and others who count hits, misses, and false positives. In this situation, the data clearly indicate wide differences in the quality of service among successful men, and since these differences are predicted by the screening procedure they have to be given a place in the evaluation system.

Goodman's discussion of the parole situation (37) exemplifies an even more comprehensive consideration of outcomes. The purpose of the prison system is to benefit society as a whole, and the parole board must take into account all the social consequences of its decision. If the man is paroled, he contributes

to the community economy through his work, his supervision by a probation officer costs so much, he perhaps commits a new crime which has both direct and indirect costs, his children are better (or worse) citizens by virtue of his presence in the home, and so on. For the man not paroled, the above outcomes arise with some altered probability when his term ends, and there is in addition the cost of his longer imprisonment to consider. To such direct outcomes must be added the indirect effect of a parole decision on the conduct of the remaining prisoners and on potential criminals in the community. Finally, the parole board gives some weight to purely human sympathy for the prisoner and his family.

It is not easy to place a value on all these consequences of a decision, especially as the effects extend indefinitely in time. In practice, it is necessary to simplify the problem. One possibility is to drop from consideration any consequences of minor importance to the institution before proceeding to more systematic evaluation. If various desirable outcomes are correlated substantially with each other, one need be little concerned to observe all outcomes since a decision rule which maximizes one of them would tend to maximize the others, and a simple weighting can account for the value of those not observed.

In many situations, however, some of the important outcomes have low intercorrelations. Amount of production per hour may be negatively related to job tenure, if there are ample opportunities for able people to find better-paying jobs. There may be a zero correlation between production and promotability, if supervision calls for different talents than routine operation. For this reason one cannot be satisfied with validating testing procedures against a single production criterion, nor with the assumption that perfect prediction of such a criterion would by itself provide the basis for an ideal selection system.

Combination of outcomes by empirical analysis

Data from multiple criteria must be reduced to a single cardinal or ordinal scale to permit decision making. That is, the various patterns of outcome must be arranged to indicate, between any two possibilities, which is preferred. Sometimes one can identify an obvious reference scale which represents the aims of the institution and can reduce other data to it. Brogden and Taylor (14) argue that in most business management the ultimate standard for judging decisions is the "dollar criterion" of the balance sheet. Some firms might deny this, insisting that preservation of certain traditions or of harmonious relations with the workers comes ahead of profit considerations. Even in these cases, however, the balance sheet provides a convenient yardstick, since there is undoubtedly some point short of bankruptcy where the firm would become concerned with income even at the sacrifice of "intangible" values, and thus the exchange rate for these intangibles can be inferred. It is not necessary that an actual criterion be accepted as primary; some abstract scale of utility may be the common metric to which all outcomes are commuted.

Value judgments are inevitable at some point in the statement of payoffs, but the number of judgments employed should be as small as possible to avoid inconsistencies in the system. Often the number of judgments can be reduced by accounting procedures, i.e., by careful empirical observation. If the sole reason for concern with spoilage, for example, is the cost of reworking or discarding the material, then a careful record of such costs indicates the seriousness of the spoilage. Brogden and Taylor describe in detail the logical and observational steps required to locate other outcomes on the "dollar criterion" scale. Military personnel decisions perhaps allow for similar accounting, since one might hope to translate certain proficiency standards into their consequent contributions to fighting power. The important outcomes of combat morale and combat efficiency are difficult to observe even in wartime, however, and cannot be studied empirically during peace. Comprehensive empirical accounting is likewise impeded by the geographical spread of military operations. It is therefore apparent that judgment will play a greater part in military than in business evaluations. The difficulties become even more overwhelming as we turn to educational and clinical institutions which seek to create long-lasting changes in people, changes of diverse kinds and with diffuse effects.

The outcomes of personnel decisions are often expressed on arbitrary scales (e.g., ratings or standard scores on proficiency measures). For evaluation it is necessary to interpret these in relation to units of production or some other absolute scale. A man's contribution depends on the quality of his performance summed over the duration of that performance. A statement in such absolute terms is required not only to take into account differences in job tenure, but also to provide an absolute estimate of benefit against which to weigh the absolute costs of the information-gathering procedure.

At first glance costs of testing and interviewing appear trivial when compared to gains in production over an extended period. More careful reasoning indicates that these costs cannot be disregarded. Experience from detailed accounting in connection with industrial inspection procedures is relevant (35). The cost of an act of inspection is very small, and the benefit from improved quality is substantial. Nonetheless, the costs of 100% inspection frequently outweigh the benefits. It was to restore this balance that sequential methods were invented, and they were regarded as having such great dollars-and-cents value that for several years they were classified as a military secret. Common sense itself provides an argument for considering costs of tests carefully. Since testing procedures do improve the quality of workers assigned to jobs, and since validity rises with length, if cost were truly negligible we would be wise to lengthen tests indefinitely. The absurdity of this proposal implies at once that there is a point of diminishing returns where costs begin to outweigh increases in quality. The best strategy for gathering information can only be determined by precise data on costs of testing.

Inferring values from decisions

The theory of utility scales has been built up from two somewhat different points of view, one concerned with eliciting direct statements of values from the decision maker, and one concerned with inferring his values from the decisions he makes. Either of these approaches can be applied to the study of personnel decisions.

The inferential approach employs the decision model in an inverse fashion. If one knows what decision a person has made in a great number of instances, then it is possible to determine what set of values is most consistent with these choices. Logically, this requires the assumption that the decision maker is following the correct strategy (except for random variation) for some set of values. A simple example of such reasoning may be taken from a study of a screening test presented elsewhere (23). This test was evaluated by its authors on the basis of the proportion of persons at each score who later succeeded or failed; the limitations of that procedure may be disregarded for the purposes of the present example. Their tabulation, after smoothing, indicates that in the recruit population the probability of success at certain score levels is as follows: at 4, 95%; at 7, 88%; at 9, 70%; at 10, 50%. Then if a given user of the test who is not subject to quota constraints applies a cutting score of 10, rejecting men with that score and above, we observe that he will accept a subgroup with a success rate of 70% but not one with 50% successes. For him, the balance point is somewhere around a success rate of 60%. Since, for him, sixty successes balance forty failures, we can say that the value of a success (relative to the zero value obtained from the rejects) is two-thirds of the cost associated with accepting a failure. For a decision maker who applies a cutting score of 9, this fraction is about one-third. He regards acceptance of a man who will fail more seriously than the decision maker who sets a higher cutting score.

This method of course does not establish what the relative values of the outcomes ought to be; it says only that the decision maker is acting as if he accepted those values. The decision maker, confronted with this information about his operations, is free to deny the appropriateness of these values. If he does, the next step is to obtain a statement of his values and translate this into a strategy which will differ from the one now in use.

A posteriori interpretation of the values which are consistent with an action sometimes throws an interesting light on decisions. Sarbin (56), in a well-known study, reported the predictions of student grade averages made by college counselors. One important finding was that these counselors consistently overestimated the students' probable grades; Sarbin regards this as a constant error or bias. This interpretation is sound if the quality of a judgment is evaluated by the absolute error of estimate. But if we hypothesize instead that the judgments are rational, we seek the evaluation system within which such "biased" judgments maximize utility. Assuming that the counselor may expect to have a symmetrically distributed error of estimate, it is sound strategy to have a mean error greater than zero if he regards overestimates as less serious than

numerically equal underestimates. It is indeed possible to calculate a set of values for over- and underestimates which would in this way rationalize any particular "constant error". It then remains to inquire whether one could reasonably regard underestimates as more serious. A counselor might contend that the underestimate has serious social consequences: it discourages the student and causes him to expect a lower standard of achievement from himself, perhaps to try less hard and to earn lower grades, perhaps even to drop out of school in view of his unsatisfactory prospects. The overestimate leads to false hopes, but the student will remain in school and obtain a valuable education even if his ultimate grades do disappoint him. Having read this casuistry (see Smith et al. (58) on the application of this term), the reader has perhaps already developed for himself a counter-argument to defend the greater seriousness in college counseling of overestimates. This is precisely the point of the example. Thousands of years of philosophy have surely established that no external judge can criticize a person's values on any grounds save internal inconsistency. If one judge prefers to emphasize one outcome and a second prefers to emphasize another, there is no scientific or logical basis to defend one emphasis over the other. The "obvious" assumption that overestimates are equal in seriousness to underestimates, i.e., that the numerical magnitude of the error is the criterion of cost, is itself based on value judgments and has no special justification over the other possibilities. This reiterates one of our fundamental theses: when mathematical formulas are used as guides to policymaking they carry hidden value judgments which the decision maker might be unwilling to accept if he considered them.

Extensive laboratory experiments have been made to infer value systems from behavior, chiefly in gambling (31, 65). These studies support the view that individuals or cultural groups have characteristic patterns of action which are consistent with different systems of values. All of these studies may be regarded as developing techniques which could be applied to important institutional decisions. At best, however, inferential procedures merely distill out values for examination. To provide a basis for future conduct, an explicit acceptance or revision of these values by the decision maker is required.

Explicit value judgments

Explicit judgment might be obtained by asking the person to state how many utility units each outcome is worth on his personal value scale. These judgments are, however, quite difficult to make and if the judgments are obtained repeatedly there will be variation in the evaluation of an outcome and inconsistency in the information about different outcomes. The most common procedure to improve the estimation of values is to employ some method of psychophysical scaling. These methods give a satisfactory degree of reliability, although it is often necessary to suppress certain inconsistencies in the data by describing them as "error" (22). Churchman and Ackoff (19) report the use of such techniques to arrive at statements of institutional values in business;

each of the examples has its parallel in personnel decisions. One firm required a quality control strategy for inspection of penicillin packages. There were eight possible types of defects, and a comparison of their seriousness was required to specify the inspection procedure so as to minimize overall risk for a given expenditure of effort. Nine persons carrying responsibility in the corporation (i.e., institutional spokesmen) participated in a scaling experiment, and from their responses a composite evaluation was established. Precisely comparable questions are involved in planning a personnel testing program where a decision must be made as to the relative importance of such various aspects of the criterion as carefulness, speed, ingenuity, stability, and leadership. A second firm described by Churchill and Ackoff was engaged in long-range planning, and used scaling methods to compare various "intangible" objectives of company policy. There is considerable similarity between this problem and that of comparing educational methods whose outcomes differ both quantitatively and qualitatively.

The problem for which comparison of outcomes seems of most immediate importance is that of classification. All schemes for differential assignment assume at some point that the outcomes for the various jobs are evaluated on the same scale. Sometimes as a stopgap all criteria are reduced to standard scores, but, as we noted in the previous chapter, this is unsatisfactory. Before such methods as those of Dwyer, Cardinet, or Votaw and Dailey can be applied, then, it will be necessary to obtain judgments of the importance of the jobs to the institution. Attention will also need to be given to the scaling of any single criterion. The assumption which is regularly made, that equal units on a criterion scale represent equal increments of value, is demonstrably false in some instances and open to question in all.

EVALUATION FOR INDIVIDUAL DECISIONS

Where a decision has a single definite outcome or configuration of outcomes, the decision maker has only to compare that result with the result of alternative decisions. No logical distinction need be made between the individual and institutional decision. Where a series of decisions is to be made, the problem becomes one of finding the correct general decision rule or strategy. The institution will evaluate each of these decisions on the same scale of values, and if that scale is one of equal intervals, the value of the set of decisions may be obtained by summation. This procedure for comparing one set of decisions to another has been the basis for all mathematical studies in the preceding chapters. In the individual decision, it is not possible to choose between courses of action save on the basis of the values of the one to whom the decision pertains. In a group of students seeking vocational counseling, the decision for each must be evaluated on a different scale of values. Since the student will make a particular choice only once, it is manifestly impossible to seek a

strategy which is superior on the average, for the average has no meaningful definition. A particular decision must be evaluated on the basis of the expected outcome and its value for this individual.

One might ultimately evaluate a decision by its actual outcome, but this appears inappropriate since factors beyond the ken of the decision maker influence the ultimate event. He might make a quite correct judgment in terms of his values and all ascertainable facts, which would nonetheless prove to be "wrong" when examined with hindsight. One's theory must make a place for the fact that the consequences of a given decision are almost never certain. Instead, with each decision is associated a certain probability distribution of expected consequences. Of two decisions, that one is preferable for which the total distribution is preferable. Since we deny the possibility of repeating the decision many times, it is meaningless to inquire which policy would lead to the greatest total utility over many applications. The judgment must be made on the distribution, not on its average.

This is a point worthy of some emphasis, because here again the transfer of statistical concepts developed for testing scientific hypotheses to decision-making has led to some misconceptions. In the literature on counseling, one finds numerous references to the responsibility of the counselor for helping the client make the right decision. The "right decision" is almost invariably interpreted as being that course of action in which his mean expectancy of success is greatest. This viewpoint has two faults. One is the implicit assumption that one particular decision is best for all persons having the same pattern of test scores. Our discussion of individual evaluations implies that two students might draw different conclusions from the same data and both be correct. Secondly, the assumption that the mean of the distribution of outcomes is the proper index must be questioned.

This problem may be considered most satisfactorily if we restrict ourselves to the simple problem of choice of curricula, where success is to be judged by grade average. Test scores might predict that a student's most probable grade average is 2.5 in curriculum A and 3.0 in curriculum B, 4.0 being the highest grade attainable. Since a choice of one curriculum may involve denial of important elements in his self-concept, or may lead toward a career which the student believes he would not enjoy, it is clear at the outset that success is unlikely to be the only criterion. When the student's values are taken into account, we may find that a 2.5 average in A will actually be more rewarding to him than a 3.0 average in B. But assume for the moment that these averages *qua* averages are equally appealing; there still remains the question of the probability distribution. Suppose that the grade average in curriculum A is difficult to predict, so that there is a large standard error of estimate. Then perhaps there is a probability of .10 of grades in the interval 3.25 - 3.75 in curriculum A, along with an equal probability of grades in the interval 1.25 - 1.75. On the other hand (to simplify), the probability may be 1.00 of a grade between 2.75 and 3.25 in curriculum B. With these alternatives, one student may prefer the

former alternative, having confidence in his ability to take advantage of the one chance in ten which would give him a very good record. Another, wishing to avoid risk, would prefer curriculum B. The choice between distributions, i.e., between hazards, rests on personality factors and affective responses. In a choice such as this, it is impossible for anyone save the decision maker himself to determine the "correct" conclusion.

This was recognized some time ago in the influential statement of Bordin and Bixler (8). They point out that the counselor, equipped with facts, can help the student recognize the probabilities attendant upon each choice. In contrast to predecessors who spoke of the counselor as telling the student the correct course of action, they left the full responsibility for choice upon the student. While their argument was based on a theory regarding the way a person learns about himself, and perhaps on a view regarding the ethics of interpersonal relations, we now supplement their reasoning with the mathematical argument that "the correct course of action" can be defined only by the person whom the decision affects.

Viewing the decision process formally as one in which information is combined with an evaluation system and knowledge of probable outcomes leads to another suggestion regarding counseling. The literature on counseling has placed great emphasis on the collecting of information and the statement of probable outcomes. However important it is to correct the client's misconceptions regarding the probable outcomes of various decisions, this is just one side of the decision process. An equally important determiner of the utility of the decision is whether the client brings to bear a consistent and fully acceptable value system. By "fully acceptable", we mean fully acceptable to himself -- consistent within itself and consistent also with his hidden, even repressed attitudes toward the world. A great deal has been written on the counseling process as a realistic self-examination, a disclosure of aptitudes and weaknesses. Clearly, it must be equally concerned with bringing the person to full awareness of his own true values.

Those who have proposed follow-up studies to evaluate the counseling process have been most concerned with the client's growth in "objective self-knowledge" (i.e., knowledge of the probable outcomes of his decisions) and with the "success" of those decisions as judged by some objective standard. But one of the main contributions of the counseling process may be to alter the person's verbalized value systems, particularly by eliminating contradictions between various wishes and hopes. It seems worthwhile therefore to study how the client's evaluation system itself changes during counseling. No one can say whether a given change is "good" or "bad", since a person is the judge of his own values. In view of the central place they play in the decision process, however, any knowledge of the nature and development of such personal evaluation systems could be helpful.

THE IMPORTANCE OF DECISION THEORY

The applied psychologist may regard test theory as but a means to his ends, a series of mathematical rules which help him to refine his procedures. Although test theory does have this function in psychology, it plays a far more important role as a sort of Gray Eminence, constantly and silently shaping the ends themselves. For test theory is both a source and an embodiment of the values which direct operations involving tests.

THE PRACTICAL POWER OF TEST THEORY

It is not too fantastic to regard test theory as exercising a control over its field comparable to that of an economic theory or a religious code over the institutions of a culture. Institutions are formed by the daily transactions of the culture, and in these transactions the participants act first of all to satisfy their needs and values. But the code which they accept brings certain values strongly to mind and thus has a deciding influence in many choices. Institutions consistent with the ostensible values receive steady encouragement, and institutions which serve other values are viewed with suspicion. Institutions for relieving needs which the code ignores can emerge only by fighting their way against the current, as it were.

A theory or code, setting forth values and assumptions explicitly, makes conformity easy. But it also displays these assumptions in full light where their inadequacies and inconsistencies can be seen and repaired. It is the task of the philosopher, economist, or political theorist to clarify issues and expose fundamental faults in the cultural system. All great historical reforms have stemmed from such revision of value assumptions. A fundamental revision of test theory will affect testing practice in the same manner that a fundamental change in political theory affects daily life in the nation that adopts it.

Test theory exercises its strongest effect by censoring new test offerings. A new clinical technique may have a dazzling debut, arousing all manner of

fluttering hopes among the young clinicians who flock around her. But off in the corner of the ballroom are the test critics, severely observing the goings-on and grimly preparing their report on the young darling's character and accomplishments. Only recently have the chaperones been armed with an "official" statement of technical recommendations by which a young test should be judged, but these standards merely formalize what has long been conveyed in books on testing, in test reviews, and in training courses. This code has made and broken reputations. While a particular test frowned on by the reviewers may attain great popularity and prosper for a time in a demi-mondaine existence, such popularity is transient. A stable place in society is awarded only to those tests which have the critics' sanction.

To speak of social and philosophical aspects of test theory, and to question the objectivity of test criticism, must astonish those who regard test theory as but a special type of mathematics. Mathematics we esteem as eternal and impeccable, one of the few embodiments of universal logic. But all mathematics rests on postulates, and while the deductions have a universal and objective character the postulates do not. Postulates are only a formal description of a situation, and in an applied problem they describe the situation as perceived by a particular observer. If we can find a satisfactory description which differs from the traditional one, mathematics permits us to erect a new theory contradicting the old at many points.

A new test theory will have numerous practical consequences. It may, for example, ultimately offer a different formula or procedure for determining which items should be retained in a test. It will certainly modify the evaluation of particular testing instruments. But what is of far greater significance is that it may channel testing effort in new directions, and open new areas of psychology to cultivation. Traditional theory has encouraged those types of tests which meet the traditional criteria, and as a result the possibilities of general intelligence tests and of achievement tests have been exploited to a high degree. Other types of tests have performed badly by the classical criteria, and psychologists have been discouraged from developing further tests of these types. At present, for example, there is great skepticism in certain circles regarding the practical usefulness of personality tests of every description, and of many types of specialized aptitude test. The development of a new variety of test drags to a halt when it proves unable to satisfy the demands of measurement theory. Any alteration in those demands may lower the esteem in which the previously successful tests are held, and raise hopes for some varieties which could not meet the old criteria. To forecast what new measurements may result from adoption of decision theory is impossible, but we may hope for such developments just as we would hope for new crops to flourish in a region where the climate changes.

DECISION THEORY COMPARED TO ALTERNATIVE MODELS

Since the beginning of the mental testing movement, the psychologist and educator have relied on one particular description of the testing process. The classic presentations of this theory are Kelley's Statistical Method (46) and Gulliksen's Theory of Mental Tests (38). These volumes offer a whole series of techniques of test construction, formulas for combining test scores, and theorems regarding the value of tests singly or in combination. Improvements in test technique during the past fifty years have been victories over defects exposed by this theory -- over subjectivity in scoring, over inadequate sampling of an ability domain, over excessive overlap between tests, etc. These important advances demonstrate the merit of this theory. Since any theory must emphasize some values at the expense of others, however, it may not give the appropriate answers to all testing problems, and our alternative theory has drawn attention to several such inadequacies.

The traditional theory views the test as a measuring instrument intended to assign accurate numerical values to some quantitative attribute of the individual. It therefore stresses, as the prime value, precision of measurement and estimation. The roots of this theory lie in surveying and astronomy, where quantitative determinations are the chief aim. In physical science, in the biometrics which was the forerunner of differential psychology, and to an increasing degree in contemporary psychological research, instruments are used for estimating quantitative variables. In pure science it is reasonable to regard the value of a measurement as proportional to its ability to reduce uncertainty about the true value of some quantity. The mean square error is a useful index of measuring power. One cannot contend that one error is more serious than another of equal magnitude when locating stars or determining melting points; measurement theory is unobjectionable when applied to such appropriate situations.

In practical testing, however, a quantitative estimate is not the real desideratum. A choice between two or more discrete treatments must be made. The tester is to allocate each person to the proper category, and accuracy of measurement is valuable only insofar as it aids in this qualitative decision. This view of testing as an aid to decision making is not unique to psychology. When a physical measuring instrument is used to control an industrial process, decision theory is more satisfactory than measurement theory for its evaluation.

Attention in traditional test theory has centered on a particular measuring instrument which is applied in the same manner to all persons. One speaks, therefore, of the validity or error of measurement of a test. When testing is seen as an aid to decision making, one recognizes that different information

may be required for different individuals. The conventional questions about "a test" do not apply to a sequential procedure. Since we cannot speak of the validity of a test which differs for every person, we must speak of the efficiency of the entire decision-making procedure.

The unsuitability of measurement theory for practical psychological testing was hinted at as long ago as 1928, when Hull (43) noted inconsistencies between measurement theory and the logic of selection problems. This discrepancy has repeatedly returned to attention, notably in the writings of Taylor and Russell and Brogden. They considered the selection problem as a choice between two alternative treatments, and arrived at conclusions inconsistent with the classical theory. Although these papers dealt with selection as an isolated problem, we may see it as one special type of personnel decision. These papers, challenging widely held beliefs as they did, may be regarded as the first ripples in a cascade of developments that follow when we consider the test in relation to the decision for which it is used.

Our formulation does not discard measurement theory. On the contrary, we have specified more accurately where it applies rigorously and where it may serve as an approximation to the ideal solution. But in every case the nature of the decision problem must be specified and these specifications must be used to determine the appropriate mathematical model. The most important parameters involved are the payoff functions for alternative treatments and the constraints upon decisions (e.g., the selection ratio). These parameters, together with the parameters of the tests (validity, reliability, intercorrelation, and cost), determine the value of the testing procedure to the decision maker.

Measurement theory has neglected the decision itself because it developed in the context of pure science. A meter stick or a thermometer is used in a particular experiment to obtain a number. Over the life of the instrument, all parts of the scale will be used, in connection with a great variety of reasoning processes. The scale is used for compound independent decisions, and it is reasonable to regard error at every point as equally serious, on the average. The psychological tester often has to decide which of several attributes to measure in a short time, but the physical scientist has not had to choose between measuring temperature and velocity or some other attribute. It is this type of choice -- which now also confronts the physicist packing measuring instruments into the limited space of an artificial satellite -- which makes it necessary to consider the importance of measurements for decision making. Measurement theory appears suitable without modification when the scale is considered in the abstract, without reference to any particular application. As soon as the scale is intended for use in a restricted context, that context influences our evaluation of the scale.

In the absence of an explicit concept of payoff for the decision maker, psychological measurement has probably been subtly influenced by quite another value consideration. It seems to us that an abstract conception of "justice" lies behind much of the concern about error of measurement. An ability test

is expected to rank persons from best to poorest, and error distorts the ranking. Since such distortion is "unfair" to the individuals who are ranked lower than they deserve, testers want to reduce error of measurement. But from a utilitarian point of view, these errors can be ignored unless they alter the goodness of whatever decisions are to be made. An attribute which is, in the abstract, deserving of reward has no bearing on allocation to treatments if it does not predict differential payoff.

A test designed to be maximally efficient for a particular decision will freely allow errors to enter if they are irrelevant to that decision. The test designed on the basis of pure measurement theory devotes testing effort to irrelevant information, and thus is not as efficient. We would expect, for example, that a test battery developed by Horst's extension of measurement theory would not be optimally efficient for classification decisions, since the theory gives no consideration to quotas and payoff functions. To date the only attempt to develop tests for a specific decision problem is the work on "peaked" tests for selection decisions. For selecting persons above a particular level of ability, it has been proposed to restrict test items to one level of difficulty. Lord (50) has recently summarized this work; the tests so designed are indeed a bit more efficient than those designed in the classical manner, but the difference is trivial in amount. Sequential procedures also sacrifice accuracy at some places in order to have greater accuracy where it most affects decisions; when costs of testing are appreciable the sequential method has clear advantages.

Just as the measurement model is not the only possible starting point for test theory, so the decision model is not the only alternative that might replace it. One prominent and rather well developed alternative is the discriminant function. The discriminant model recognizes that discrete alternatives may confront the tester, regarding his basic task as the allocation of each individual to the proper one of several categories. It differs from our approach chiefly in the way payoffs are specified. Each person is said to "belong" in a particular treatment, and the evaluation of a strategy consists of some suitably weighted function of the number of erroneous assignments. Sometimes this way of formulating a problem will be ideal, although it is then possible to regard it as a special case within decision theory. Another evaluation is, we believe, more often suitable in personnel classification. We may consider payoff within a treatment as varying continuously with score, and allocate people so as to maximize total payoff. This avoids any assumption that persons fall into homogeneous types, or that payoff is the same for all "correct" decisions. To put it differently, the discriminant function is concerned entirely with between-groups variance in test score, whereas our model takes into account the fact that predictable variance in payoff within groups dictates information-gathering and assignment strategies.

Another alternative to traditional measurement theory stresses the fact that test data are qualitative rather than truly quantitative. Ferguson, Guttman, Loevinger, and others view the test as an instrument for locating persons in

categories. This theory recognizes that the overall validity coefficient of a test is a less important quality than its power to discriminate at the boundary lines where decisions are to be made. Each author has proposed an index of the goodness of a test as a categorizing instrument, and these indices could be used in place of the error of estimate or the validity coefficient. Test construction could point toward maximization of such indices. Again, the decision model provides a vantage point for inspecting the proposal. In a discussion of the Ferguson-Thurlow approach reported elsewhere (23), we show that this method of describing the testing process has two defects. It treats all errors of discrimination as equally serious, even when the order of categories makes some errors far more serious than others. Secondly, it compares the decisions made by the test to chance decisions, instead of to the best a priori strategy. Since the decision model permits us to use the strategy and payoff matrix required by a given practical problem, it is more flexible than these alternative proposals, but will lead to the same conclusions in the limited situations where their assumptions apply.

As compared with the alternative theories, decision theory has one distressing characteristic. Instead of the definite formulas and procedures for test construction which other theories offer, decision theory is a general model for stating any particular testing problem. With certain commonplace assumptions we can generally derive formulas expressing the value of a test and principles for designing efficient tests. The formulas, however, involve so large a number of parameters that they are difficult to comprehend, and many of them can be evaluated only by numerical integration. As compared with the algebra of measurement theory or the discriminant function, the mathematics of decision theory is involved and laborious. This is the price paid for bringing in the parameters required to describe a problem rigorously. Where this price is too high, the tester can make simplifying assumptions to obtain approximate answers to his questions. Decision theory states the questions comprehensively, and thereby reduces the likelihood that an inadequate approximation will be taken as an exact and final answer.

THE VALUE OF GENERAL ABILITY TESTS

If test theories gild or blemish reputations, the clearest way to see the consequences of a new theory is to apply it to particular types of tests. The application does more than improve our understanding of the test discussed. It makes the important concepts of the theory more meaningful, and spotlights the factual and theoretical questions to which research must be directed. We begin by discussing the general ability test, which of all tests is the most successful by the standards of measurement theory; in the subsequent section, we turn to some of the procedures such as the interview and projective techniques which have been especial targets of criticism.

By general ability tests, we think particularly of measures of scholastic or intellectual ability of the type widely used in predicting school and job performance. Such tests are general because the attribute they measure is related to success in a great number of undertakings.

In the prevalent theory, the validity coefficient is the cachet of test nobility. The coefficient of determination looks scornfully on any test which cannot account for twenty per cent or more of criterion variance, and the coefficient of forecasting efficiency sets even more haughty standards. The general ability test is virtually the only test whose coefficients consistently come within the range from .40 to .80, and this fact has made it the most universal of procedures for psychological assessment. No new theory alters the empirical fact that this type of instrument has high technical quality; does decision theory, however, modify the way the instrument is judged?

For some tests, perhaps, taking into account the importance of a decision as we do by means of σ_c would cause us to lower our estimate of the value of the test. But the criteria on which the general ability test bears are often of the highest importance, and indeed are closely related to the person's success and adjustment in his entire life. Furthermore, decision theory argues that a test which bears on many decisions makes far more contribution than the test which bears on only one, and this is where the general test excels. A single IQ for a high school student tells teachers of a dozen subjects whether or not to press for higher accomplishment, tells what occupational range the student may properly aim toward, helps in decisions about college education, and so on without end. From this point of view, the traditional theory seems if anything too restrained in its enthusiastic appraisal of the general ability test.

Other aspects of decision theory, however, temper this enthusiasm. We need not dwell on the finding previously summarized (pp. 74-77) that utility is most often proportional to validity, and that the coefficient of forecasting efficiency sets indefensible standards. While this result does alter the way we judge a test of validity .40 as compared to one of validity .20 or .30, the change is rather minute as compared to other consequences of our theory.

The argument that a testing strategy must be compared to the best a priori strategy applies with particular force to general tests. If valid decisions can be made using information already at hand or cheaply obtainable, the test should be judged by the increase in validity it offers. The zero-order validity coefficient is relevant only if, without the test, the decision maker would be forced into chance decisions. Just because the general test deals with qualities required in a wide range of performances, information on those qualities can be obtained from past performances. We are told by Professor P. E. Vernon that in England success in grammar school has been predicted from properly scaled primary-school ratings with validity coefficients as high as .80 (for the total range of talent). In a typical study, the best test battery raised the multiple correlation by less than .05. A validity coefficient of .82

for a general ability test makes it appear spectacularly valuable, but its true contribution to decision making is scarcely a five per cent improvement over non-test decisions. In view of the importance of the decision, even this gain very likely repays the effort of testing. (But see our further discussion below.) When the estimated value of the test is deflated to proper proportions, however, one immediately wonders whether tests of rather low validity, bearing on decisions where substitute data are not available, might have greater social importance. It is only speculative to suggest that such importance might be found in predictors of delinquency, emotional disorders, or vocational interests; for a test to be valuable as a screening device, it must not only predict with some accuracy, but there must be beneficial treatments available for those singled out by the test.

In academic decisions, treatments in which payoff is closely related to test score are certainly available. Yet that very success of the general ability test as a predictor may imply that the tests are not right for the uses to which they are put. This question was first raised by the realization that in counseling of college students the most common problem is a choice between curricula or vocations. Since the general test predicts grades in every curriculum to about the same degree, it sheds no light on the decision of the typical student. For this reason, differential tests of abilities and interests have been sought which would predict between-treatment differences. The same argument has been applied to military classification, with a resultant increase in use of differential predictors.

Decision theory has formalized the problem of payoff in differential decisions, and extended it in two ways. First, it has been shown that in classification the value of measuring a dimension depends in a complex manner on the payoff functions relating outcome to the various dimensions and on the quotas for the treatments (13). These relations have to date been studied only superficially, but the results tend to devalue the general test. Second, a previously unrecognized type of decision has been identified, namely the placement decision. Placement decisions are a sub-species of classification decisions and the same principles apply. Since only one score-continuum is involved, the factors affecting utility can be thoroughly traced. Moreover, this special case covers many of the important applications of tests.

A placement decision is involved when persons are grouped on the basis of their scores in a general ability test, and each group receives a different treatment. Now for each treatment there is a relation between payoff and score, which for simplicity we assume to be linear. The slope is determined by the variation of payoff σ_e and the validity coefficient r_{ye} for that treatment. The contribution of the test to a placement decision depends on the differences between slopes for the several treatments, and therefore the value of a test for sectioning students will be zero if the slope of the payoff function is the same for all treatments. A general test is, by its very nature, positively correlated with success in a great variety of treatments, and its value for placement is limited by that fact.

The correlation of .85, more or less, between prediction and grammar-school success in England is far less significant than it seems at first glance, because a very large correlation would surely be found also in the "modern secondary" school to which most of the rejectees go. From the viewpoint of national policy, the aim is to maximize total output. The present scheme does this only if the slope $\sigma_{e r_{ye}}$ is greater for grammar schools than for modern secondary schools. The facts to determine this are not available, simply because it has previously been thought that predictive validity alone is required. Very likely, σ_e for an unselected range of talent is greater in grammar schools than in modern secondary schools, and if so the allocation procedure would be profitable. But the profit is less than the $\bar{r}$ of .85 suggests. Moreover, better bases for decision could be invented if differential payoff were held in mind as the goal. The general ability predicts success in both treatments. There must be attributes (methods of problem solving? preference for abstract thought? character traits? interests?) which are more relevant to one mode of instruction than the other. The slopes of payoff functions for these qualities will be less steep than for the general test, but the difference in slope from treatment to treatment and the benefit from testing will be greater.

British secondary school allocation has afforded a useful example of a placement problem, but it is not unique. Nearly all the everyday uses of ability and achievement tests in American schools are for placement rather than selection decisions; our observations apply there as well. And the observation which applies with greatest force is that the basic data required to determine the worth of placement tests have never been obtained.

A decision model has forced us to formulate one more concept which bears on the reputation of ability tests, and that is the notion of adaptive treatment. American testers have placed themselves at the disposal of institutions wishing to assign men to predetermined treatments, and have pointed with pride to the fact that they often can raise the average output by careful selection. But nearly equal gains can often be attained by another branch of personnel technology which refuses to accept the treatment as given. The job-simplification expert and the human engineer seek to fit the job to unselected men. The greater their success the less the value of selection. The tester has failed to realize that he is competing with the treatment simplifier. And the latter's method is often the more economical, for his changes may be made permanent while the tester must evaluate new employees forever.

The true problem is to find the optimum combination of adaptation of treatment and selection of persons. This calls, not for rivalry between engineering psychologist and tester, but for intimate cooperation. As we have stated the problem abstractly, one can attain a given goal (e.g., training of radar operators) by an infinite variety of treatments which vary continuously over many dimensions. For every person there is a best treatment, and for every treatment a best type of person. Only by a study of the payoff surface linking treatment dimensions and payoff dimensions can one arrive at the best combination. This calls for a combination of the experimental methods of the job designer

and the differential-correlational methods of the tester. The vast problem thus posed is impossible to cast in concrete terms at present, but if we look at just one aptitude at a time, we can make some progress.

For any one type of score such as general ability we have a family of straight lines relating score to payoff for various treatments which in its simplest form is characterized by its envelope, a second-degree curve (Figure 9). And the greater the curvature of this line, the greater the value of adaptation. If the curvature is great, adaptation to unselected men gains much of the value testers have been claiming for selection. Adaptation may be accomplished by routinization, reduction of the verbal components of training, and other means. Even in schools and universities, elimination of intellectual hurdles which are irrelevant to the person's later performance could reduce the importance of selection on general intellectual characteristics. There are practical limits to modification of treatment, but these limits are astonishingly remote.

Putting all the foregoing argument together, we can neither exalt nor debase the general ability test. Considering the importance of decisions and the compound uses of general information introduces favorable factors. The best a priori strategy, with or without adaptation, is often much better than a chance decision, and neglect of this has caused the test to be overrated. For placement decisions, the test has surely received much too much credit; it may even have been used where its true value is negligible.

This analysis makes clear that the question whether a given test is profitable or unprofitable is meaningless. The value of a test is great for one decision and small even for other decisions involving the same treatment. A generalized study of payoff functions for a type of test can tell us far more than we now know regarding its usefulness. But it has no universal value. Its contribution must ultimately be judged in terms of the particular decision problem the decision maker has before him.

THE VALUE OF WIDE-BAND PROCEDURES

Of all the implications of decision theory, perhaps the most dramatic is the new interpretation it offers for wideband procedures. In counseling and guidance, one must ordinarily help the person answer several questions at once, and each answer involves somewhat different types of information. The counselor may obtain information narrowly focussed on one dimension, or may use a procedure which covers many areas. Bandwidth, i.e., greater coverage, is purchased at the price of lowered fidelity or dependability. For any decision problem there is an optimum bandwidth. This conclusion departs from conventional theory, which assumes that it is always desirable to maximize dependability.

Among the important wideband procedures are the interview, the projective technique, the essay examination, and analysis of patterns of successes and

failures on ability tests. Each of the wideband devices is unsatisfactory, by the usual standards of predictive efficiency and reliability. Our work suggests, however, that the negative evidence may not bear on the usefulness of the procedures for the function they best fulfill.

Let us begin with the interview, a technique which is universally accepted by those who make personnel decisions. No matter how elaborate a testing program may be, it is supplemented by an interview wherever the cost of interviewing permits. Indeed, some of the very psychologists who report negligible validity for interview judgments themselves insist on interviewing the prospects when they have a vacancy to fill in their own offices. The evidence, however, is preponderantly negative. Interviewers' judgments disagree with each other and with objective criteria. Why, in the face of such evidence, does the interview retain its popularity?

If the interview were omitted, estimates of intelligence and other qualities which can be judged by objective means would be more valid than when contaminated by interviewers' impressions. By omitting the interview, however, the employer would give up the possibility of obtaining information on characteristics which formal measurement procedures do not reveal. An interview permits the decision maker to cover the length and breadth of the subject's history and character. "How did you get along with your last employer? What aspects of your work interested you most? Tell me about your family. . . ." In this varied conversation, significant facts can come to light which no structured test or questionnaire could reveal. A reluctance to talk about a particular job experience perhaps discloses a failure. Other remarks hint at a competitive attitude which would be an aid in one job and a hazard in another. The personality traits which receive chief attention as one interview develops may be brushed over in an interview with another person. The virtue of the interview is that it can turn in any of hundreds of directions, following leads in a way that the structured narrowband procedure cannot.

Quite similar comments could be made regarding the projective technique, which touches on abilities, interests, social relations, sexual attitudes, methods of thinking, and so on through hundreds of characteristics. Any given record says next-to-nothing about some of these areas, but each record presents a few striking individual features of undoubted importance to the clinician. The essay test has a smaller range. As a measure of factual knowledge a history paper compares more or less unfavorably with an objective test in the same area. The essay test, however, gives information on spelling, organization of ideas, manner of attack on a problem, imagination and originality, special biases, and so on. The teacher scoring such a paper has a better chance to "get to know" the student than does the teacher scoring an objective test.

The wideband procedure sheds light on many different decisions. The employment interview does not bear only on whether to hire the person. It suggests how to assign him, how to supervise him, what weaknesses must be allowed for, what are his prospects for promotion, and so on. The projective

test helps the clinician to decide regarding dozens of steps of treatment. And the essay test may provide numerous suggestions about educational procedures.

One may contrast two different aspects of any information source: exhaustiveness and dependability. Exhaustiveness is the extent to which the information offered covers the ground we wish to know about. Dependability is the extent to which the information offered is true. Only a precise evaluation of a given procedure can indicate whether it strikes the proper balance between the two desirable qualities. Though any one fact or judgment from a wideband procedure is undependable, the procedure as a whole contributes more in some circumstances than a narrowband procedure. The wideband procedure can perhaps improve every decision that is to be made. The narrowband procedure gives accurate information with respect to one decision, but provides no guidance at all for the remaining decisions, leaving them to be made on a priori grounds.

Even where a wideband technique does contribute more than a precise narrowband instrument, the decision maker should be reluctant to rely on untrustworthy information. Usually it is practical to make the wideband instrument the first stage in a sequential process, arriving at reversible rather than terminal decisions. A sequential process makes ideal use of fallible data, trusting them only to the extent that they deserve. Actually, in most present practice, wideband devices are used sequentially rather than for final decisions.

The interview is used to identify questions which need to be considered and facts which need to be obtained. In vocational guidance, the client mentions activities which have interested him, and the counselor notes that all of them involve detailed stereotyped work. This leads, first, to further questions about interests to see whether the hypothesis of preference for stereotyped activities is sound. Secondly, it suggests the possibility of underlying emotional insecurity which can be investigated by intensive interviewing, a personality test, or information from acquaintances. For another client, the salient observation is a strong artistic interest coupled with poor grades in art courses. The next steps are to verify the genuineness of the interest by questions as to what about art he likes, to determine the reasons for his poor grades by questioning his instructors, and perhaps to apply tests of artistic talent. In counseling of a student referred for low grades, the interview would suggest whether the critical questions relate to lack of interest in his curriculum, poor study methods, an undesirable attitude, poor reading, an emotional problem, or something else. Dependable measuring instruments exist for several of these areas, and would form the next stage in fact-finding. It would be possible to describe in similar detail the sequential nature of employment interviewing, clinical diagnosis, or educational diagnosis.

In the ideal situation, the decision maker comes to no conclusion on the basis of the wideband procedure. It directs his subsequent observations, and perhaps suggests a tentative treatment which will enhance his opportunity to observe what is pertinent. Ultimately, he hopes to obtain enough data for a highly dependable terminal decision.

The merit of the broad survey as a first step in decision making is obvious. In a hundred counseling interviews attention is directed to a hundred different questions, each of which appears salient in only a fraction of the cases. All this ground could not possibly be covered by systematic, precise techniques. The survey is fallible; it misses some points, and opens up false leads. But the subsequent questioning reveals the false leads as false, and there are fewer missed points than there would be if a narrowband procedure had been used.

The wideband procedures, though condemned on the basis of validation research, have won a large and loyal following who claim to have had satisfactory experience with the methods. We submit that the enthusiasm of their users is the result of satisfactory experience with them in their sequential application. The critic, on the other hand, has most often applied standards appropriate for evaluating terminal decisions. The interview gives an undependable impression of intelligence. Judgments of aggression based on the Rorschach disagree too often with observed social behavior. Psychiatrists interviewing the same person frequently arrive at quite different estimates of emotional stability. Qualitative analysis of Binet or Wechsler protocols is a poor way to measure flexibility or creativeness. The evidence is that estimates from wideband techniques are rather unreliable and have low validity. Used sequentially, however, they draw attention to facts which would otherwise be missed. The price of pursuing false leads is one which the decision maker is often willing to pay.

In recent years controversy has centered on projective methods, which have been an active subject of research. The proponents of these methods have tended to reject the negative evidence, claiming that the techniques permit highly valid judgment. There is no objective support for the claim that projective data are a valid basis for terminal decisions. Supporters of projective tests have seemingly rested their case on the wrong grounds. We suggest that they have argued for the conventional validity of their tests because that has been the only passport to respectability in test theory. Decision theory points out the existence of quite another virtue in tests and the available evidence is not inconsistent with the claim that projectives have this virtue.

We believe that proponents and critics of projective tests could agree on the following statements: Inferences from projective tests are often wrong, but right more often than chance would allow. Projective techniques frequently suggest hypotheses about the individual which, when they are confirmed, have great practical importance. Many of the suggestions and hypotheses about idiosyncratic characteristics of the individual are subsequently proved baseless. The theory and technique of projective testing and interpretation can be greatly improved.

Reconciliation of the objective evidence with the clinician's enthusiasm is made possible by the fact that the evidence disproves claims which should never have been made, but supports claims consistent with much clinical practice. Our argument should be compared with Meehl's discussion (53) of the value of clinical judgment in diagnosis. His conclusion that clinical methods

are chiefly useful for constructing hypotheses to be confirmed by further investigation is strikingly similar to ours.

Placing wideband techniques in a logical framework encouraging to their advocates does not absolve them from all criticism. There is some risk that our clarification of logical possibilities will be taken as an endorsement of present practices. Nothing could be farther from our intentions. Projective and other wideband methods are used in indefensible ways which should be criticized and corrected. We hope that our theoretical analysis directs criticism toward the essential points and away from false issues.

Recognizing the much greater value of wideband techniques when used sequentially will, we hope, discourage the use of these methods for terminal decisions. Research on these methods should consider their usefulness as a way of focussing investigation. It is necessary to determine how often the leads suggested by a given wideband procedure are fruitful. Considering this along with the importance of the findings and the cost of the procedure is the only legitimate basis for evaluating the method. A low "validity" is acceptable if the cost of separating the true leads from the false is tolerable.

There is a continuum from great to narrow bandwidth. The tester can spread himself too far, just as he can confine himself too narrowly. By structuring the interview or the projective technique one can reduce the bandwidth with a corresponding gain in fidelity. The fact that the bandwidth of current projective methods is almost unlimited suggests that it would be profitable to restrict them to some degree. In the TAT, for example, better information about relations with parents would presumably be obtained by increasing the proportion of plates involving parent figures. In many diagnostic problems the attendant sacrifice of information about other aspects of the person would surely be justified.

The true issue is not whether wideband methods are good or bad. The question is, for any given decision problem what is the best information-gathering procedure? If a projective technique would be useful, just how should it be designed and interpreted to give the greatest help? Nothing in our argument leads us to think that any procedure can conceivably give an accurate analysis of "the whole personality". Instead, it is necessary to distill from a limited quantity of information the most intelligent possible decision. The problem is to find the procedure which, in the time available, offers the greatest yield of important, relevant, and interpretable information.

A FINAL WORD

Our search for a coherent theory of decision-making on the basis of test information has led in many directions. Wherever we turn, we find unanswered questions. Research of many types is needed, ranging from simple fact finding to major inventions. Mathematical studies of test design and allocation strategies, new types of empirical studies of payoff surfaces and of validity, and

even philosophical studies of evaluation problems have a part to play in extending test theory. Along with these theoretical developments we perceive many possibilities of replacing present measuring instruments with new ones: placement tests of superior differential validity, personality surveys with appropriate bandwidth for particular tasks, collections of brief tests for sequential classification, and so on. Our final conclusion may be optimistic or pessimistic, but it is inescapable. The test theory developed to date covers only a small corner of the domain within which the decision maker operates, and in only a few specialized instances have we discovered the testing practices which will most increase the utility of personnel decisions.

UTILITY IN FIXED-TREATMENT SELECTION AND PLACEMENT

Definition of the selection problem

The following development formally defines the selection problem where information is univariate and the treatment to be given accepted men is specified independently of the test. The argument indicating the relation of utility to test validity is a restatement of that presented by Brogden (10, 12), and Cochran (20). We make the following assumptions:

1. Decisions are made regarding an indefinitely large population of persons. This "a priori population" consists of all applicants after screening by any procedure which is presently in use and will continue to be used.
2. Regarding any person, i , there are two possible alternative decisions: accept (t_A) and reject (t_B).
3. Each person has a test score y_i , which has zero mean and unit standard deviation.
4. For every person there is a payoff e_{it_A} which results when the person is accepted. This payoff has a linear regression on test score. The test will be scored so that r_{ye} is positive.
5. When a person is rejected, the payoff e_{it_B} results. This payoff is unrelated to test score, and may be set equal to zero.
6. The average cost of testing a person on test y is C_y , where $C_y > 0$.
7. The strategy will be to accept high scoring men in preference to others. A cutoff y' will be located on the y continuum so that any desired proportion $\phi(y')$ of the group falls above y' . Above that point probability of acceptance is 1.00; below it, zero. Cochran shows that such a strategy is optimal for selection with fixed quota.

From assumption (4), the expected payoff from accepting a man is

$$e_{yt_A} = \sigma_e r_{ye} y + e_{ot_A} \quad (1.1)$$

Here, σ_e and r_{ye} are computed on the a priori population, and e_{ot_A} is the payoff expected when an average man in that population is assigned to treatment t_A .

Utility as a function of validity in selection

The expected utility of a set of decisions is obtained by summing e_{yt} , the expected payoff, over all persons and subtracting cost of testing. Let $\underline{U}$ represent the total utility resulting from decisions about $\underline{N}$ persons, cost of testing being taken into account.

$$U = \sum_{i=1}^N (p_{t_A/y_i} e_{y_i t_A} + p_{t_B/y_i} e_{y_i t_B} - C_y) \quad (1.2)$$

Using assumption (5) and (1.1),

$$U = \sum_{i=1}^N p_{t_A/y_i} (\sigma_e r_{ye} y_i + e_{ot_A}) - NC_y \quad (1.3)$$

If individuals in any y array are accepted, their total utility is

$$U_y = N p_y p_{t_A/y_i} (\sigma_e r_{ye} y + e_{ot_A}) - NC_y \quad (1.4)$$

From assumption (7), $\int_{y'}^{\infty} p_y dy = N \phi(y')$. Therefore, summing over all values of y ,

$$U = N \sigma_e r_{ye} \int_{y'}^{\infty} p_y y dy + N \phi(y') e_{ot_A} - NC_y \quad (1.5)$$

The a priori utility U_o is the utility resulting when $N \phi(y')$ men are selected from the population by chance:

$$U_o = N \phi(y') e_{ot_A} \quad (1.6)$$

The gain in utility from test y is

$$\Delta U = U - U_o = N \sigma_e r_{ye} \int_{y'}^{\infty} p_y y dy - NC_y \quad (1.7)$$

It will be noted that, for any a priori distribution, gain in utility is a linear function of r_{ye} . We shall be chiefly concerned with the case where we can assume a normal distribution of y . We shall let $\xi(y)$ and $\phi(y)$ represent the

ordinate and area (upper tail) corresponding to any value of y under this assumption. Then (1.7) becomes

$$\Delta U = N\sigma_e r_{ye} \xi(y') - NC_y \quad (1.8)$$

which is the net gain from testing N men. Hereafter, we shall divide (1.8) by N and let ΔU represent the gain in utility per man tested.

As we proceed to study a variety of types of decision, we shall employ B_{yd} as a general symbol for the contribution of a particular test y to a specified decision d , apart from cost of testing. In fixed-treatment selection, $B_{yd} = \sigma_e r_{ye} \xi(y')$, so that the gain in utility per man tested

$$\Delta U = B_{yd} - C_y \quad (1.9)$$

Rewriting (1.8) in this form will simplify several later mathematical developments.

The gain in utility per man accepted is

$$\sigma_e r_{ye} \frac{\xi(y')}{\phi(y')} - \frac{C_y}{\phi(y')} \quad \text{or} \quad \sigma_e r_{ye} \bar{y} - \frac{C_y}{\phi(y')}, \quad (1.10)$$

since the average test score $\bar{y}$ for accepted men equals $\xi(y')/\phi(y')$.

Region where test is admissible. From (1.8) it is evident that $\Delta U > 0$ for any y' such that

$$\xi(y') > \frac{C_y}{\sigma_e r_{ye}} \quad (1.11)$$

For any positive value of $C/\sigma_e r$ there are two values of y' , symmetrically located about $y = 0$, beyond which selection with the test is less beneficial than random selection. Outside these limiting values, use of the test is an inadmissible strategy, in Wald's sense.

Optimum strategy. The maximum utility per person accepted is obtained by differentiating (1.10) with respect to $\phi(y')$:

$$\sigma_e r_{ye} \frac{y' \phi(y') - \xi(y')}{\phi^2(y')} + \frac{C_y}{\phi^2(y')} = 0 \quad (1.12)$$

$$\xi(y') - y' \phi(y') = \frac{C_y}{\sigma_e r_{ye}} \quad (1.13)$$

Only one value of y' satisfies this equation. Equation (1.13) is meaningful only when $\xi(y') > C/\sigma_e r$, since testing is otherwise not profitable. Therefore

$y' \phi(y)$ must be positive, and y' must be greater than zero.

When there is a fixed numerical quota $N \phi(y')$, setting the optimum y' according to (1.13) determines the optimum number of persons to be tested. For maximum utility with a fixed quota, enough men should be tested so that the cutting point is above the mean.

Utility in placement with fixed quotas

The above relationship may be generalized to consider placement decisions where any number of fixed treatments are to be used.

For each of the n treatments there is a linear payoff function, whose parameters will vary from treatment to treatment. The y continuum is to be divided into n segments, each containing some predetermined proportion of the cases. Each segment is bounded by two cutting scores, y_t' and y_t'' , the latter being the upper boundary. y_t'' is also, of course, the lower boundary for the next treatment, $t+1$. $\phi(y_t')$ - $\phi(y_t'')$ is the quota for treatment t .

If all men above y_t' were assigned to treatment t the utility per man tested would be (from (1.5), invoking the normal assumption),

$$\sigma_{e_t} r_{ye_t} \xi(y_t') + \phi(y_t') e_{ot} - C_y \quad (1.14)$$

A similar expression may be written in terms of y_t'' . The utility when men between y_t' and y_t'' are so assigned is, by subtraction,

$$U_t = \sigma_{e_t} r_{ye_t} [\xi(y_t') - \xi(y_t'')] + e_{ot} [\phi(y_t') - \phi(y_t'')] - C_y \quad (1.15)$$

We may write $\Delta \xi_t$ and $\Delta \phi_t$ for the bracketed terms, so that

$$U_t = \sigma_{e_t} r_{ye_t} \Delta \xi_t + e_{ot} \Delta \phi_t - C_y \quad (1.16)$$

Over all treatments,

$$U = \sum_t \sigma_{e_t} r_{ye_t} \Delta \xi_t + \sum_t e_{ot} \Delta \phi_t - C_y \quad (1.17)$$

The a priori strategy with fixed treatment is to assign randomly selected men to each treatment to fill the quota. From (1.6), for each treatment

$$U_{ot} = e_{ot} \Delta \phi_t \quad (1.18)$$

Subtracting,

$$\Delta U = \sum_t \sigma_{e_t}^r y_{e_t} \Delta \xi_t - C_y \quad (1.19)$$

This again may be written as $B_{yd} - C_y$, B_{yd} thus being defined in fixed-treatment placement by the first term of (1.19).

Utility in placement with adaptive quotas

If treatments are fixed and arranged according to the slopes of their payoff functions, but quotas are not specified, it is possible to increase utility by altering the quotas. Maximizing (1.17) with respect to y_t^1 , and using $t-1$ and t to indicate two adjacent treatments such that $y_{t-1}'' = y_t^1$,

$$\begin{aligned} \frac{\partial U}{\partial y_t^1} &= \sigma_{e_{t-1}}^r y_{e_{t-1}} y_t^1 \xi(y_t^1) + e_{o(t-1)} \xi(y_t^1) \\ &\quad - \sigma_{e_t}^r y_{e_t} y_t^1 \xi(y_t^1) - e_{ot} \xi(y_t^1) = 0 \end{aligned} \quad (1.20)$$

Simplifying,

$$\sigma_{e_{t-1}}^r y_{e_{t-1}} y_t^1 + e_{o(t-1)} = \sigma_{e_t}^r y_{e_t} y_t^1 + e_{ot} \quad (1.21)$$

or

$$y_t^1 = \frac{e_{o(t-1)} - e_{ot}}{\sigma_{e_t}^r y_{e_t} - \sigma_{e_{t-1}}^r y_{e_{t-1}}} \quad (1.22)$$

Equation (1.22) specifies the best dividing point between the two treatments. Similar boundaries may be set between any other adjacent pair of treatments. Each value of y^1 locates a point at which the payoff functions for two treatments intersect. If such computations for any treatment yield a value of $y_t'' < y_t^1$, that treatment drops out of the series of treatments used. Its quota is zero.

When the decision maker is allowed to alter quotas, his best a priori strategy is to assign all men to whichever of the allowed treatments gives greatest utility when $y = 0$. We will designate it t_o . Considering all treatments, if adaptive quotas are used,

$$\Delta U = \sum_t \sigma_{e_t}^r y_{e_t} \Delta \xi_t + \sum_t (e_{ot} - e_{ot_o}) \Delta \phi_t - C_y \quad (1.23)$$

A PAYOFF SURFACE DEPENDING ON A UNIVARIATE APTITUDE

To consider problems involving choice among many treatments, we develop a general function relating payoff to a univariate aptitude. Parameters are employed such that this surface will fit a variety of actual data.

It was pointed out in Chapter III that when many scores are available it is possible to perform a factor analysis, deriving the orthogonal factors $s_j = s_1, s_2, \dots$. Of these factors, the ones which determine payoff under any of the treatments are referred to as aptitude factors. In this appendix we restrict ourselves to the case where all tests under consideration measure only one aptitude dimension $\underline{s}$ (although they may also measure common factors not related to payoff). We assume that $\underline{s}$ has zero mean, unit standard deviation, and normal distribution.

We assume that the expected payoff under any treatment is a linear function of $\underline{s}$:

$$e_{st} = \sigma_{e_t} r_{se_t} s + e_{ot} \quad (2.1)$$

Our problem is to examine the possible changes in payoff when treatments are allowed to vary. As discussed in Chapter III, treatments may vary along many dimensions. As the treatment varies, either σ_{e_t} or r_{se_t} or both may vary, and e_{ot} may vary. Change in either of the first two parameters affects the slope of the payoff function, and we therefore introduce a slope parameter $m_{st} = \sigma_{e_t} r_{se_t}$. If two treatments have the same slope, it will always be preferable to assign men to the treatment for which e_{ot} is greater. For each slope m_{st} , then, only one of the set of treatments need be considered.

We desire to postulate a relatively simple form for the payoff surface and therefore assume that the intercepts e_{ot} vary in the following manner:

$$e_{ot} = c + bm_{st} - am_{st}^2, \text{ where } a > 0 \quad (2.2)$$

This assumption is adopted because it is the simplest function which defines a non-trivial case. While $\underline{b}$ is likely to be greater than zero, this is not required. Substituting in (2.1), we have the equation for the surface relating expected payoff to aptitude $\underline{s}$:

$$e_{st} = m_{st}s + c + bm_{st} - am_{st}^2 \quad (2.3)$$

The parameters $\underline{a}$, $\underline{b}$, and $\underline{c}$ would change if a different aptitude were measured, or if a different set of treatments were under consideration. Empirical values for the parameters can only be obtained by administering several of the treatments in the series to persons for whom an estimate of s_i is available.

At any level of aptitude $\underline{s}$, there is a treatment t_s which yields the greatest payoff. Maximizing $\underline{e}$ with respect to m_{st} ,

$$\frac{\partial e}{\partial m_{st}} = s + b - 2am_{st} = 0 \quad (2.4)$$

$$m_{st_s} = \frac{s + b}{2a} \quad (2.5)$$

Substituting this in (2.3) gives the maximum payoff obtainable for a person having aptitude $\underline{s}$:

$$e_{st_s} = \frac{(s + b)^2}{4a} + c \quad (2.6)$$

When a group with average ability $\underline{s}$ is assigned to the treatment best for the average man in the group, the utility per man assigned is expressed by (2.6).

appendix 3

UTILITY WITH ADAPTIVE TREATMENT

It will be advantageous to determine the utility from placement with adaptive treatment before considering adaptive selection. The opposite order is employed in describing the findings in the text.

Placement with fixed quotas

It is assumed that a large number of treatments are available, and that a particular test is under consideration. The aptitude $\underline{s}$ is common to test and treatments, and the payoff function is described by (2.3). The decision maker is to divide persons into a certain number of groups, the proportion to be placed in each group being specified. Each group will receive a different treatment, the decision maker being allowed to select the treatment best fitted to the estimated aptitude of each group. Throughout this discussion, utility will refer to utility per man tested; that is to say, $\underline{N}$ is dropped from all equations.

If the test is not given, the best a priori strategy under these conditions is to divide men at random into groups, and to assign all groups to the treatment best when $s = 0$. From (2.5), this treatment t_o is such that

$$m_{t_o} = b/2a \quad (3.1)$$

This is not the same t_o as appeared in (1.23). If $\phi(y')$ persons, randomly selected, are assigned to t_o , it follows from (2.6) that the a priori utility per man tested is

$$U_{ot_o} = \phi(y')(c + b^2/4a) \quad (3.2)$$

When the test y is used as a basis for dividing the group, utility is increased. As in Appendix 1, we assume that the cutting scores y' and y'' are fixed so that $\Delta\phi_t = \phi(y'_t) - \phi(y''_t)$ is the quota for treatment $\underline{t}$.

The average test score of men between y' and y'' is $\frac{\xi(y'_t) - \xi(y''_t)}{\Delta\phi_t}$ or $\frac{\Delta\xi_t}{\Delta\phi_t}$. Assuming a linear regression of $\underline{s}$ on $\underline{y}$, the average value of $\underline{s}$ in this group is $r_{ys} \frac{\Delta\xi_t}{\Delta\phi_t}$. From (2.6), the payoff when these individuals are given the treatment optimum for them is,

$$e_{st_s} = \frac{\Delta\phi_t}{4a} (r_{ys}^2 \frac{\Delta^2\xi_t}{\Delta^2\phi_t} + b)^2 + c\Delta\phi_t \quad (3.3)$$

The utility is

$$U_t = \frac{1}{4a} (r_{ys}^2 \frac{\Delta^2\xi_t}{\Delta\phi_t} + 2br_{ys}\Delta\xi_t + b^2\Delta\phi_t) + c\Delta\phi_t - \Delta\phi_t C_y \quad (3.4)$$

Over all treatments, $\sum_t \Delta\xi_t = 0$ and $\sum_t \Delta\phi_t = 1$. Therefore

$$U = \frac{r_{ys}^2}{4a} \sum_t \frac{\Delta^2\xi_t}{\Delta\phi_t} + \frac{b^2}{4a} + c - C_y \quad (3.5)$$

From (3.2) and (3.5) the gain in utility is

$$\Delta U = \frac{r_{ys}^2}{4a} \sum_t \frac{\Delta^2\xi_t}{\Delta\phi_t} - C_y \quad (3.6)$$

In adaptive placement where quotas are specified, therefore, ΔU is a function of r_{ys}^2 -- regardless of quotas or number of treatments. B_{yd} is here equivalent to $\frac{r_{ys}^2}{4a} \sum_t \frac{\Delta^2\xi_t}{\Delta^2\phi_t}$.

For men assigned to any one treatment, the gain in utility is found by subtracting (3.2) from (3.4).

$$\Delta U_t = \frac{r_{ys}^2}{4a} \frac{\Delta^2\xi_t}{\Delta\phi_t} + \frac{r_{ys}b\Delta\xi_t}{2a} - \Delta\phi_t C_y \quad (3.7)$$

Adaptive selection

The expression for utility in adaptive selection now follows readily from (3.7). In adaptive selection a quota to be accepted, $\phi(y')$, is specified, and all accepted men are assigned to the one most suitable treatment. The remainder are rejected, and their payoff is zero.

A priori, a randomly selected group of size $1 - \phi(y')$ would be rejected, and payoff is zero. After testing, the utility is zero less the cost of testing

the rejected persons. Therefore the gain in utility from testing depends entirely on the increase in utility for accepted men, less the cost of testing the rejectees. Modifying (3.7) to take into account the fact that $y_t'' = \infty$, and subtracting $(1 - \phi(y^1))C_y$, the gain in utility is

$$\Delta U = \frac{r_{ys}^2}{4a} \frac{\xi^2(y^1)}{\phi(y^1)} + \frac{r_{ys} b \xi(y^1)}{2a} - C_y \quad (3.8)$$

The first two terms of (3.8) define B_{yd} for adaptive selection.

To clarify certain points discussed in the text, expressions are required for utilities which might result under several different conditions. Consider three treatments: t_A , some treatment which happens to be in use a priori; t_o , which is best suited to randomly selected men; and t_s , suited to men with test scores above y^1 . U_{ot} is the utility from assigning $\phi(y^1)$ randomly selected men to treatment t . From (2.2):

$$U_{ot_A} = \phi(y^1)(c + b m_{t_A} - a m_{t_A}^2) \quad (3.9)$$

U_{ot_o} has already been given in (3.2). Substituting (2.5) in (2.2) with $s = o$, $r_{ys} = \frac{\xi(y^1)}{\phi(y^1)}$, we have

$$U_{ot_s} = \phi(y^1) \left[c + \frac{b^2}{4a} - \frac{1}{4a} \left(r_{ys} \frac{\xi(y^1)}{\phi(y^1)} \right)^2 \right] \quad (3.10)$$

The a posteriori utilities for men selected on the basis of test y may be denoted U_{yt} . The average ability for these men is $r_{ys} = \frac{\xi(y^1)}{\phi(y^1)}$, hence

$$U_{yt_A} = \phi(y^1) \left[m_{t_A} r_{ys} \frac{\xi(y^1)}{\phi(y^1)} + c + b m_{t_A} - a m_{t_A}^2 \right] - C_y \quad (3.11)$$

$$U_{yt_o} = \phi(y^1) \left[\frac{b r_{ys} \xi(y^1)}{2a \phi(y^1)} + \frac{b^2}{4a} + c \right] - C_y \quad (3.12)$$

$$U_{yt_s} = \phi(y^1) \left[\frac{1}{4a} \left(r_{ys} \frac{\xi(y^1)}{\phi(y^1)} + b \right)^2 + c \right] - C_y \quad (3.13)$$

Placement with adjusted quotas

We shall next consider the case where the maximum number of treatments is fixed, but the decision maker may alter the quotas as well as the treatments. The optimum cutting score y_t^1 between any two treatments $t-1$ and t may be located by differentiating (3.6) with respect to y_t^1 .

$$\frac{\partial \Delta U}{\partial y_t^1} = \frac{r_{ys}^2}{4a} \left[\frac{2\Delta\phi_{t-1}\Delta\xi_{t-1}y_t^1\xi(y_t^1) - \Delta^2\xi_{t-1}\xi(y_t^1)}{\Delta^2\phi_{t-1}} - \frac{2\Delta\phi_t\Delta\xi_t y_t^1\xi(y_t^1) - \Delta^2\xi_t\xi(y_t^1)}{\Delta^2\phi_t} \right] = 0 \quad (3.14)$$

Simplifying this expression,

$$y_t^1 = \frac{1}{2} \left[\frac{\Delta\xi_t}{\Delta\phi_t} + \frac{\Delta\xi_{t-1}}{\Delta\phi_{t-1}} \right] \quad (3.15)$$

This states that the boundary between adjacent treatments should ideally be located halfway between the midpoints for the intervals.

This equation applies to any cutting point. With n_t treatments, there are $n-1$ boundaries, and (3.15) provides $n-1$ simultaneous equations. No general solution for this set of equations has been found.

appendix **4**

TWO-STAGE SEQUENTIAL STRATEGY

Two tests, y_a and y_b , are available. One of these, designated y_1 , will be given to all persons. Persons scoring above a certain value y_1' will be accepted at once, and persons scoring below some other value y_1'' will be rejected. The remaining persons will be given the second test and will be finally accepted or rejected on the basis of the combined information from both tests. The appropriate cutting scores y_1' and y_1'' to provide the desired proportion ϕ_A of accepted men are to be determined.

Men are to be accepted for a fixed treatment t_A . The assumptions of Appendix 1 will be retained, save for the introduction of a second test. Since the two tests y_a and y_b may be correlated, y_2 is used to designate the standardized component of the designated second test which is independent of the first. The validities of the two independent scores are r_{y_1e} and r_{y_2e} . These two scores may be combined into a battery score $\underline{Y}$, using weights which maximize the battery validity r_{Ye} .

$$r_{Ye}^2 = r_{y_1e}^2 + r_{y_2e}^2 \quad (4.1)$$

$$Y = \frac{r_{y_1e}y_1 + r_{y_2e}y_2}{r_{Ye}} \quad (4.2)$$

Since all persons for whom $\underline{Y}$ is constant contribute equal expected payoff (from 1.1), the optimum cutoff for selecting among the men given the second test is a value $Y = Y'$ (shown as line MN in Figure 11). We now seek the values of Y' , y_1' , and y_1'' which give the greatest utility per man tested, with any selection ratio.

Men in Region II of Figure 11 are accepted on the basis of the first test. The gain in utility for these men, over the utility from men chosen by chance, using (1.8) and momentarily disregarding cost, is

$$\Delta U_{II} = \sigma_e r_{y_1} e^{\xi(y_1'')} \quad (4.3)$$

In Region III, the analogous gain is

$$\Delta U_{III} = \int_{y_1=y_1'}^{y_1''} \int_{Y=Y'}^{\infty} \sigma_e r_{Y} e^{\xi(y_1, y_2)} dy_1 dy_2 \quad (4.4)$$

Men in regions I and IV are rejected and thus their utility is zero. Let $y_{2.1}$ be the cutting score on y_2 for a specified y_1 , and also let y_2' and y_2'' represent the values of $y_{2.1}$ corresponding to y_1' and y_1'' . From (4.2),

$$y_{2.1} = \frac{r_{Y} e^{Y'} - r_{y_1} e^{y_1}}{r_{y_2} e} \quad (4.5)$$

Noting that $\xi(y_1, y_2) = \xi(y_1) \xi(y_2)$, (4.4) becomes

$$\begin{aligned} \Delta U &= \sigma_e \int_{y_1=y_1'}^{y_1''} \int_{y_{2.1}}^{\infty} (r_{y_1} e^{y_1} + r_{y_2} e^{y_2}) \xi(y_1) \xi(y_2) dy_1 dy_2 \\ &= \sigma_e \int_{y_1'}^{y_1''} [r_{y_1} e^{y_1} \phi(y_{2.1}) + r_{y_2} e^{\xi(y_{2.1})}] \xi(y_1) dy_1 \end{aligned} \quad (4.6)$$

We may designate the costs of the two tests as C_{y_1} and C_{y_2} . Since all persons are given the first test, while those between y_1' and y_1'' are given the second, the average cost per man tested is

$$C_{y_1} + C_{y_2} [\phi(y_1') - \phi(y_1'')] \quad (4.7)$$

The net gain in utility as a result of testing,

$$\begin{aligned} \Delta U &= \sigma_e \xi(y_1'') r_{y_1} e + \sigma_e \int_{y_1'}^{y_1''} [r_{y_1} e^{y_1} \phi(y_{2.1}) + r_{y_2} e^{\xi(y_{2.1})}] \xi(y_1) dy_1 \\ &\quad - C_{y_1} - C_{y_2} [\phi(y_1') - \phi(y_1'')] \end{aligned} \quad (4.8)$$

The selection ratio is the combined proportion in II and III.

$$\phi_A = \int_{y_1''}^{\infty} \xi(y_1) dy_1 + \int_{y_1'}^{y_1''} \xi(y_1) \phi(y_{2,1}) dy_1 \quad (4.9)$$

Since $\phi(y_{2,1})$ is a function of Y and y_1 , equation (4.9) establishes a relationship between y_1' , y_1'' , and Y' . Therefore only two of these can be independent variables. To maximize ΔU we follow the method of Lagrange for a restrained maximum, introducing an undetermined multiplier λ . Three simultaneous equations define the conditions for the maximum:

$$\frac{\partial \Delta U}{\partial y_1'} + \lambda \frac{\partial \phi}{\partial y_1'} = 0$$

$$\frac{\partial \Delta U}{\partial y_1''} + \lambda \frac{\partial \phi}{\partial y_1''} = 0 \quad (4.10)$$

$$\frac{\partial \Delta U}{\partial Y'} + \lambda \frac{\partial \phi}{\partial Y'} = 0$$

Substituting the appropriate values in (4.10) and transposing,

$$\begin{aligned} \sigma_e r_{y_1} e^{y_1'} + \frac{\sigma_e r_{y_2} e^{\xi(y_2')} - C_{y_2}}{\phi(y_2')} &= -\lambda \\ \sigma_e r_{y_1} e^{y_1''} - \frac{\sigma_e r_{y_2} e^{\xi(y_2'')} - C_{y_2}}{1 - \phi(y_2'')} &= -\lambda \\ \sigma_e r_{Y} e^{Y'} &= -\lambda \end{aligned} \quad (4.11)$$

Eliminating λ from equations (4.11) and noting that $1 - \phi(x) = \phi(-x)$ and that $\xi(x) = \xi(-x)$ gives

$$\begin{aligned} r_{y_1} e^{y_1'} + r_{y_2} e^{\frac{\xi(y_2')}{\phi(y_2')}} - \frac{C_{y_2}}{\sigma_e \phi(y_2')} &= r_Y e^{Y'} \\ r_{y_1} e^{y_1''} - r_{y_2} e^{\frac{\xi(-y_2'')}{\phi(-y_2'')}} + \frac{C_{y_2}}{\sigma_e \phi(-y_2'')} &= r_Y e^{Y'} \end{aligned} \quad (4.12)$$

These equations define y_1' and y_1'' respectively in terms of $r_{y_1} e$, $r_{y_2} e$, C_{y_2}/σ_e , and Y' .

For computational purposes it is advantageous to determine the y_2 coordinates of points $\underline{M}$ and $\underline{N}$, i.e., y_2' and y_2'' . We use equation (4.5) to eliminate Y' from (4.12),

$$r_{y_1} e^{y_1'} + r_{y_2} e^{\frac{\xi(y_2')}{\phi(y_2')}} - \frac{C_{y_2}}{\sigma_e \phi(y_2')} = r_{y_1} e^{y_1'} + r_{y_2} e^{y_2'} \quad (4.13)$$

$$r_{y_1} e^{y_1''} - r_{y_2} e^{\frac{\xi(-y_2'')}{\phi(-y_2'')}} + \frac{C_{y_2}}{\sigma_e \phi(-y_2'')} = r_{y_1} e^{y_1''} + r_{y_2} e^{y_2''}$$

Hence

$$y_2' = \frac{\xi(y_2')}{\phi(y_2')} - \frac{C_{y_2}}{\sigma_e r_{y_2} e^{\phi(y_2')}} = -y_2'' \quad (4.14)$$

Thus y_2' and y_2'' are equal in absolute value and depend only on $C_{y_2}/\sigma_e r_{y_2} e$.

The optimum pre-reject or pre-accept strategy could be derived separately. However, the fact that the partial derivative with respect to y_1' is independent of y_1'' in (4.11) shows that the values of y_1' and y_1'' so obtained would be identical to those for the complete sequential strategy.

In order to simplify comparison of the various testing strategies it is advantageous to express (4.12) and (4.14) in terms of the correlation of each test with the total battery rather than in terms of their validity coefficients. The validity of either stage of testing can be written thus:

$$r_{y_1} e = r_{y_1} Y^r Y_e \text{ and } r_{y_2} e = r_{y_2} Y^r Y_e \quad (4.15)$$

The terms $r_{y_1} Y_1$ and $r_{y_2} Y$ are not independent since

$$r_{y_1}^2 Y + r_{y_2}^2 Y = 1 \quad (4.16)$$

Then (4.12) becomes

$$r_{y_1} Y y_1' + r_{y_2} Y \frac{\xi(y_2')}{\phi(y_2')} - \frac{C_{y_2}}{\sigma_e r_{Y_e} \phi(y_2')} = Y' \quad (4.17)$$

$$r_{y_1} Y y_1'' - r_{y_2} Y \frac{\xi(-y_2'')}{\phi(-y_2'')} + \frac{C_{y_2}}{\sigma_e r_{Y_e} \phi(-y_2'')} = Y'$$

and for (4.14) we have

$$y_2' = \frac{\xi(y_2')}{\phi(y_2')} - \frac{C_{y_2}}{\sigma_e r_{Ye} r_{y_2 Y} \phi(y_2')} = -y_2'' \quad (4.18)$$

Using this formulation it is possible to treat $C_{y_2}/\sigma_e r_{Ye}$ as a single constant C_2 which is given (or can be specified hypothetically). Figure A-1 based on (4.18) permits ready determination of y_2' .

To solve a particular problem one begins with $r_{y_1 Y}$ and C_2 , and some trial value of Y' such that $\phi(Y')$ is near the desired selection ratio. Then y_2' and y_2'' are obtained from Figure A-1, and y_1' and y_1'' are found from (4.5).

Any Y' yields one pair of first-screen cutoffs. The selection ratio obtained for the trial Y' is found from (4.9). The gain in utility is given by (4.8), but it is convenient to rewrite that equation in this form:

$$\begin{aligned} \frac{\Delta U}{\sigma_e r_{Ye}} = & r_{y_1 Y} \xi(y_1'') + \int_{y_1'}^{y_1''} \left[r_{y_1 Y} y_1 \phi(y_{2.1}) + r_{y_2 Y} \xi(y_{2.1}) \right] \xi(y_1) dy_1 \\ & - \frac{C_{y_1}}{\sigma_e r_{Ye}} - \frac{C_{y_2}}{\sigma_e r_{Ye}} \left[\phi(y_1') - \phi(y_1'') \right] \quad (4.19) \end{aligned}$$

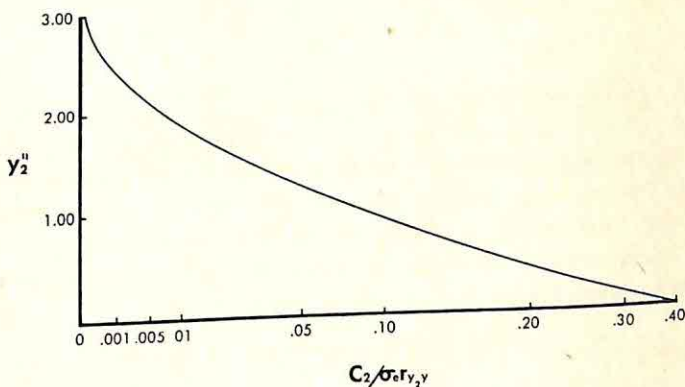


Figure A-1. Chart for determining optimal cutting scores on second test

There is no simple procedure for solving directly from ϕ_A . Successive trials with interpolation will determine the values of y_1' , y_1'' , and Y' that yield a desired ϕ_A .

appendix 5

OPTIMUM LENGTH OF A TEST

A test is said to be lengthened homogeneously when we combine unit tests having the same cost, equal intercorrelations $r_{y_1 y_1}$, and the same relation to criteria (i.e., the same $r_{y_1 e}$). The relation of the unit test to the criterion is assumed to be the same regardless of the unit's position in the series. Correlation increases with length k according to this well-known equation:

$$r_{y_k e_t} = r_{y_1 e_t} \sqrt{\frac{r_{y_k y_k}}{r_{y_1 y_1}}} = r_{y_1 e_t} \sqrt{\frac{k}{1 + (k-1)r_{y_1 y_1}}} \quad (5.1)$$

A similar equation may be written for r_{y_s} .

The cost, formerly denoted by C_y , may be divided into two components, C_0 and C_1 . C_0 consists of the basic cost of setting up the test, giving directions, etc.; C_0 is constant as k changes. C_1 represents the costs associated with taking a test of unit length (supplies, time of examiner and subjects if paid, etc.). This cost increases proportionally to length. Hence for a test k times as long as the unit test, the total cost

$$C_{y_k} = C_0 + kC_1 \quad (5.2)$$

Fixed treatment, fixed quotas

From (1.19), gain in utility per man tested in either selection or placement is

$$\Delta U_k = \sum_t \sigma_{e_t} r_{y_k e_t} \Delta \xi_t - C_y = B_{y_k d} - C_{y_k} \quad (5.3)$$

Substituting from (5.1) and (5.2),

$$\Delta U_k = B_{y_1 d} \sqrt{\frac{r_{y_k y_k}}{r_{y_1 y_1}}} - (C_0 + k C_1) \quad (5.4)$$

$B_{y_1 d}$ is defined for fixed treatment by the value of the first term of (5.3) when $\underline{k}$ is 1. The function has a maximum, provided $\Delta U > 0$ for some value of $\underline{k}$. The maximum occurs where

$$\frac{\partial \Delta U}{\partial k} = \frac{B_{y_1 d} (1 - r_{y_1 y_1})}{2 \sqrt{k} [1 + (k-1) r_{y_1 y_1}]^{3/2}} - C_1 = 0 \quad (5.5)$$

This condition may be rewritten thus:

$$\left[(1 - r_{y_1 y_1}) \frac{B_{y_1 d}}{2 C_1} \right]^2 = k [1 + (k-1) r_{y_1 y_1}]^3 \quad (5.6)$$

One root of this equation gives the desired optimum length.

In selection or placement a single-stage battery of $\underline{v}$ tests may be used. If $R_{Y_{v e_t}}$ is the correlation of the battery score with payoff,

$$\Delta U = \sum_t \sigma_{e_t} R_{Y_{v e_t}} \Delta \xi_t - C_y \quad (5.7)$$

Since utility of the battery with the $\underline{v}$ th test removed is expressed by the same formula with $\underline{v}-1$ substituted for $\underline{v}$, the gain from adding the $\underline{v}$ th test is the increment

$$\Delta U_v = \sum_t \sigma_{e_t} \Delta \xi_t (R_{Y_{v e_t}} - R_{Y_{v-1 e_t}}) - C_v \quad (5.8)$$

$R_{Y_{v e_t}} - R_{Y_{v-1 e_t}}$ is unlikely to be the same for all treatments.

If the problem is one of selection for a single treatment, (5.8) becomes

$$\Delta U = (R_{Y_{v e_t}} - R_{Y_{v-1 e_t}}) \sigma_{e_t} \xi_t - C_v = \Delta R_v \sigma_{e_t} \xi_t - C_v \quad (5.9)$$

We may now consider the advisability of adding a unit test of such length that its variable cost is C_1 . The battery may be augmented profitably by adding a unit of any test already in the battery if for that unit

$$\Delta R_v > \frac{C_1}{\sigma_{e_t} \xi_t} \quad (5.10)$$

or a unit of any test y not yet in the battery if for that unit

$$\Delta R_v > \frac{C_{0y} + C_1}{\sigma_{e\xi_t}} \quad (5.11)$$

Among all units available, that one should be added first which provides the greatest increment in utility. Horst (39) has developed general procedures for adjusting the length of the tests in a battery so as to maximize R_{Ye} , under the assumption that C_0 is negligible.

Adaptive placement, fixed quotas

When treatments may be adjusted to fit the aptitude s of selected men, from (3.6)

$$\Delta U_k = \frac{r_{y_k s}^2}{4a} \sum_t \frac{\Delta^2 \xi_t}{\Delta \phi_t} - C_{y_k} \quad (5.12)$$

Substituting from (5.1) and (5.2),

$$\Delta U_k = B_{y_1 d} \frac{r_{y_k y_k}}{r_{y_1 y_1}} - (C_0 + kC_1) \quad (5.13)$$

Here $B_{y_1 d}$ is the benefit under adaptive treatment, defined by the first term of (5.12) when $k = 1$. This function has a maximum (provided $\Delta U > 0$ for any k) where

$$\frac{\partial \Delta U}{\partial k} = B_{y_1 d} \frac{(1 - r_{y_1 y_1})}{[1 + (k-1)r_{y_1 y_1}]^2} - C_1 = 0 \quad (5.14)$$

This condition may be rewritten:

$$(1 - r_{y_1 y_1}) \frac{B_{y_1 d}}{C_1} = [1 + (k-1)r_{y_1 y_1}]^2 \quad (5.15)$$

The positive root is the desired optimum length k . Notation being simplified,

$$k = 1 - \frac{1}{r} + \frac{1}{r} \sqrt{\frac{B(1-r)}{C}} \quad (5.16)$$

Adaptive selection, fixed quotas

In selection, from (3.8)

$$\Delta U_k = \frac{1}{4a} \frac{\xi^2(y')}{\phi(y')} r_{y_k}^2 + \frac{b}{2a} \xi(y') r_{y_k} - C_{y_k} \quad (5.17)$$

Hence

$$\Delta U_k = \frac{1}{4a} \frac{\xi^2(y')}{\phi(y')} r_{y_1}^2 \left[\frac{r_{y_k} y_k}{r_{y_1} y_1} \right] + \frac{b}{2a} \xi(y') r_{y_1} \sqrt{\frac{r_{y_k} y_k}{r_{y_1} y_1}} - (C_0 + kC_1) \quad (5.18)$$

$$\begin{aligned} \frac{\partial \Delta U_k}{\partial k} &= \frac{1}{4a} \frac{\xi^2(y')}{\phi(y')} r_{y_1}^2 \frac{(1 - r_{y_1} y_1)}{\left[1 + (k-1) r_{y_1} y_1\right]^2} \\ &\quad + \frac{b}{2a} \xi(y') r_{y_1} \frac{(1 - r_{y_1} y_1)}{2\sqrt{k} \left[1 + (k-1) r_{y_1} y_1\right]^{3/2}} - C_1 \quad (5.19) \end{aligned}$$

Setting this equal to zero gives an expression in k one of whose roots is the optimum length of test for adaptive selection.

STRATEGY FOR MULTI-STAGE SELECTION

The following appendix describes the procedures developed by the Statistical Research Group (SRG) (35, 62) for dividing persons into two groups so as to minimize risks associated with misclassification. We have paraphrased their argument to fit the personnel problem.

A unit test y_1 is to be used at each stage of testing. The person has a "true score" y_∞ which is his average score after an infinite number of stages of testing. Since y_∞ is a predictor of criterion performance, it is desired to select persons whose true quality is above a certain level y'_∞ , either rejecting the remainder or assigning them to a second treatment. The reliability $r_{y_1 y_\infty} = r_{y_1 y_\infty}^2$ is known; assuming the variance of true scores to be 1.00, the standard deviation σ of y_1 for persons having the same y_∞ can be computed.

While no serious loss attaches to a decision to accept a person slightly below y'_∞ , or to reject a person slightly above this value, losses become more serious when erroneous decisions are made about persons farther from the cutting line. In the usual manner for Neyman-Pearson risks, two values $y'_\infty - w_1$ and $y'_\infty + w_2$ are located. A judgment is then made as to the tolerable risk α of accepting a person whose true ability is $y'_\infty - w_1$, and the tolerable risk β of rejecting a person whose true ability is $y'_\infty + w_2$. These risks constitute the specifications for a desirable strategy.

Strategy. Three functions are now calculated:

$$h_1 = b\sigma^2/(w_2 - w_1) \quad (6.1)$$

$$h_2 = a\sigma^2/(w_2 - w_1) \quad (6.2)$$

$$S = (w_1 + w_2)/2 \quad (6.3)$$

The parameters $\underline{a}$ and $\underline{b}$ are functions of α and β which have been tabled by SRG:

$$a = \ln(1 - \beta)/\alpha \quad b = \ln(1 - \alpha)/\beta \quad (6.4)$$

The person's score is totalled at the end of every stage of testing. If, at the k th stage, the score is greater than $h_2 + kS$, the person is accepted; if less than $h_1 + kS$, the person is rejected. If it is between these values, testing is continued. A simple graphical procedure for making these decisions is available.

Operating characteristic and selection ratio. The strategy has an operating characteristic function, which describes P_{A/y_∞} , that is, the probability of accepting a person at each true quality level. This function is approximately ogival in shape and passes through the points (w_1, α) and $(w_2, 1 - \beta)$, and is described by

$$P_{A/y_\infty} = (e^{g_1} - 1)/(e^{g_2} - 1) \quad (6.5)$$

Here, e is the base of the natural logarithms and the parameters g_1 and g_2 are defined by

$$g_1 = 2h_1(S - y_\infty)/\sigma^2 \quad g_2 = 2(h_1 + h_2)(S - y_\infty)/\sigma^2 \quad (6.6)$$

If, as in our work, the distribution of y_∞ is assumed to be known, the $\Sigma p_{y_\infty} P_{A/y_\infty}$ indicates the proportion of applicants who will be accepted, or the selection ratio.

Average sample number and cost of testing. The number of tests to be given each person will depend on his scores. For a large number of persons having the same true score, we can predict the average number of stages of testing k which will be required.

$$k_{y_\infty} = (h_2 + h_1)P_{A/y_\infty} + \frac{h_1}{S - y_\infty} \quad (6.7)$$

Multiplying by the probability of each y_∞ , and summing, gives the total number of tests. Introducing the cost of the unit test then permits an estimate of the total cost of all tests.

Calculating operations. A computer program has been prepared which permits calculation of the relation between selection ratio, average quality of persons accepted, and cost of testing. The program assumes that $\alpha = \beta$, which is not a very restrictive assumption, and that ability is normally distributed and error of measurement uniform for all values of y_∞ . The parameters for any given calculation are α , $r_{y_1 y_1}$, and C_1 , the cost per unit test. A working parameter y'_∞ is also introduced, but drops out of the functions obtained. This program was employed to obtain the results given in the text.

The program also permits a person seeking a good strategy for a particular decision to start with a known selection ratio and cost of testing. Assuming a value for $w_1 = w_2$ and using nomograms not reproduced here, one can determine appropriate values of α and y_{∞}^1 for these conditions, and then derive the strategy from (6.1) and (6.2). This strategy will probably be different from the true optimum obtained by Wald's recursion formula, but it is expected to be a good approximation.

DISTRIBUTION OF EFFORT IN COMPOUND DECISIONS

When several decisions are to be made, and a test is available relevant to each, a decision must be made regarding the optimum length of each test. That is to say, the composition of the testing battery is to be adjusted so as to maximize utility. Assume that there are w independent decisions, each dependent on a different aptitude so that payoffs are uncorrelated. There is a test y_d bearing on decision d and irrelevant to the others. The unit of length for each test is such that its cost is C_1 . The remaining parameters for each test are $B_{y_1 d}$, $r_{y_1 y_1 d}$, and C_{0d} . We shall assume that C_{0d} depends only on the time necessary to distribute the test, give instructions, etc., and thus is fixed for each test regardless of its length but can vary from test to test.

For any test, as in (5.2) the cost at length k is

$$C_{y_k d} = C_{0d} + kC_1 \quad (7.1)$$

We shall fix the total allowable cost C_T for the battery. Since we assume that costs depend only on time, this is equivalent to assuming a fixed total testing time.

Fixed treatment

As in (5.4), if treatments and quotas are fixed, when test y_d is used at length k_d , the gain in utility per man for that decision is

$$\Delta U_d = B_{y_1 d} \sqrt{\frac{r_{y_k y_k d}}{r_{y_1 y_1 d}}} - (C_{0d} + k_d C_1) \quad (7.2)$$

For the w decisions, there are w tests, but we shall consider the utility of any battery of v tests where $v \leq w$. Then summing with respect to d ,

$$\sum_1^v C_{y_k d} = \sum C_{0d} + C_1 \sum k_d = C_T \quad (7.3)$$

$$\sum_1^v \Delta U_d = \sum B_{y_1 d} \sqrt{\frac{r_{y_k y_k d}}{r_{y_1 y_1 d}}} - C_T \quad (7.4)$$

It may be noted from (7.3) that the total number of items ($\sum k_d = K$) which can be used for any battery is a constant determined by $(C_T - \sum C_{0d})/C_1$. In order to maximize $\sum \Delta U_d$, it is necessary again to use the method of Lagrange as in (4.10). With respect to any k_d , as in (5.5),

$$\frac{\partial \sum \Delta U_d}{\partial k_d} = \frac{B_{y_1 d} (1 - r_{y_1 y_1 d})}{2\sqrt{k_d} [1 + (k_d - 1)r_{y_1 y_1}]^{\frac{3}{2}}} - C_1 = 0 \quad (7.5)$$

Letting λ be an unspecified multiplier, and retaining the constraint of (7.3), we obtain a series of v equations of the form

$$\frac{B_{y_1 d} (1 - r_{y_1 y_1 d})}{2\sqrt{k_d} [1 + (k_d - 1)r_{y_1 y_1}]^{\frac{3}{2}}} - C_1 + \lambda C_1 = 0 \quad (7.6)$$

No algebraic solution for this system of equations has been obtained.

Special cases. We can treat the equation directly when $v = 2$. Then, from (7.3),

$$K = \frac{C_T - C_{01} - C_{02}}{C_1} \quad (7.7)$$

$$\frac{B_{y_1 1} (1 - r_{y_1 y_1 1})}{2\sqrt{k_1} [1 + (k_1 - 1)r_{y_1 y_1 1}]^{\frac{3}{2}}} = \frac{B_{y_1 2} (1 - r_{y_1 y_1 2})}{2\sqrt{K - k_1} [1 + (K - k_1 - 1)r_{y_1 y_1 2}]^{\frac{3}{2}}} \quad (7.8)$$

This equation involving the single unknown k_1 indicates the optimum distribution of effort between two tests, when other parameters are fixed.

A second special case to be considered is that where all v tests and the related decisions are uniform, i.e., $r_{y_1 y_1 d}$, $B_{y_1 d}$, and C_{0d} are the same for all decisions. Then, from (7.6), k_d is the same for every decision. From (7.3),

$$v = \frac{C_T}{C_0 + kC_1} \quad (7.9)$$

This permits us to inquire as to the optimum number of tests to be used. (7.4) becomes

$$\sum \Delta U_d = v B_{y_1 d} \sqrt{\frac{k}{1 + (k-1)r_{y_1 y_1}}} - C_T \quad (7.10)$$

Substituting for $\underline{v}$ (using (7.9)), and maximizing ΔU with respect to $\underline{k}$,

$$\begin{aligned} \frac{\partial \Delta U}{\partial k} = 0 = & \frac{C_T B_{y_1 d} (1 - r_{y_1 y_1})}{2\sqrt{k}(C_0 + kC_1) [1 + (k-1)r_{y_1 y_1}]^{3/2}} \\ & - \sqrt{\frac{k}{1 + (k-1)r_{y_1 y_1}}} \frac{C_T C_1 B_{y_1 d}}{(C_0 + kC_1)^2} \end{aligned} \quad (7.11)$$

Simplifying,

$$(1 - r_{y_1 y_1})(C_0 + kC_1) - 2kC_1 [1 + (k-1)r_{y_1 y_1}] = 0 \quad (7.12)$$

$$k = \frac{1 - r_{y_1 y_1}}{4r_{y_1 y_1}} \left[-1 + \sqrt{1 + 8 \frac{C_0 r_{y_1 y_1}}{C_1 (1 - r_{y_1 y_1})}} \right] \quad (7.13)$$

This value of $\underline{k}$ yields maximum utility. The second root of (7.13) is ignored because it yields a negative value of $\underline{k}$, which has no practical meaning. The optimum $\underline{v}$ is determined by substitution in (7.9).

Adaptive treatment

A similar analysis for adaptive placement leads to a simpler solution. The equation analogous to (7.4) is, from (5.13),

$$\sum_1^v \Delta U_d = \sum_1^v B_{y_1 d} \frac{r_{y_k y_k d}}{r_{y_1 y_1 d}} - \sum_1^v (C_{0d} + k_d C_1) \quad (7.14)$$

The method of Lagrange leads to $\underline{v}$ equations of this form:

$$B_{y_1 d} \frac{(1 - r_{y_1 y_1 d})}{[1 + (k-1)r_{y_1 y_1 d}]^2} - C_1 + \lambda C_1 = 0 \quad (7.15)$$

Letting $\mu = \frac{1}{1 - \lambda}$, we find that

$$k_d = 1 - \frac{1}{r_{y_1 y_1 d}} + \frac{\mu}{r_{y_1 y_1 d}} \sqrt{\frac{B_{y_1 d}(1 - r_{y_1 y_1 d})}{C_1}} \quad (7.16)$$

These equations can be solved for $\underline{d}$ for any set of tests. If we let

$$D_{y_1 d} = \frac{1}{r_{y_1 y_1 d}} \sqrt{\frac{B_{y_1 d}(1 - r_{y_1 y_1 d})}{C_1}} \quad (7.17)$$

and recall that $\Sigma k_d = K$ is a constant determined by test costs, we find that

$$k_d = 1 - \frac{1}{r_{y_1 y_1 d}} + \mu D_{y_1 d} \quad (7.18)$$

$$\text{and} \quad \Sigma k_d = K = v - \sum_1^v \frac{1}{r_{y_1 y_1 d}} + \sum_1^v \mu D_{y_1 d} \quad (7.19)$$

$$\text{Therefore} \quad \mu = \frac{1}{\sum_1^v D_{y_1 d}} \left[\sum_1^v \frac{1}{r_{y_1 y_1 d}} - v + K \right] \quad (7.20)$$

Denoting any particular $\underline{d}$ by d' ,

$$k_{d'} = 1 - \frac{1}{r_{y_1 y_1 d'}} + \frac{D_{y_1 d'}}{\sum_1^v D_{y_1 d}} \left[\sum_1^v \frac{1}{r_{y_1 y_1 d}} - v + K \right] \quad (7.21)$$

There is one such equation for each of the tests.

These equations may be used to compute the optimum distribution of effort. With $\underline{w}$ decisions, there are many possible subsets of tests. These equations give a distribution of effort for any set, but do not indicate directly which set is best. Various possible subsets may be compared by calculating $\Sigma \Delta U_d$ after lengths are optimally adjusted, to determine which test is most profitable.

Special cases. If all decisions are uniform, from (7.14),

$$\Sigma \Delta U_d = v B_{y_d} \frac{r_{y_k y_k}}{r_{y_1 y_1}} - C_T \quad (7.22)$$

Substituting for $\underline{y}$ and $r_{y_k y_k}$, and differentiating,

$$\frac{\partial \Sigma \Delta U_k}{\partial k} = \left[\frac{C_T}{C_0 + kC_1} \right] \frac{B_{y_1} d (1 - r_{y_1 y_1})}{[1 + (k-1)r_{y_1 y_1}]^2} - \frac{k B_{y_1} d (C_T C_1)}{[1 + (k-1)r_{y_1 y_1}][C_0 + kC_1]^2} = 0 \quad (7.23)$$

$$\frac{1 - r_{y_1 y_1}}{1 + (k-1)r_{y_1 y_1}} - \frac{kC_1}{C_0 + kC_1} = 0 \quad (7.24)$$

$$k = \sqrt{\frac{C_0(1 - r_{y_1 y_1})}{C_1 r_{y_1 y_1}}} \quad (7.25)$$

This is the optimum length for each test in uniform adaptive placement decisions.

BIBLIOGRAPHY

1. Adams, E. W. A survey of Bernoullian utilities and applications. New York: Bureau of Applied Social Research, Columbia University, 1954.
2. Anderson, G. L. Quantitative thinking as developed under connectionist and field theories of learning. In Esther J. Swenson, G. L. Anderson, and C. L. Stacey, Learning theory in school situations. Minneapolis: University of Minnesota Press, 1949. Pp. 40-73.
3. Arbous, A. G., and Sichel, H. S. On the economies of a pre-screening technique for aptitude test batteries. Psychometrika, 1952, 17, 331-346.
4. Arrow, K. J. Social choice and individual values. New York: Wiley, 1951.
5. Berkson, Joseph. "Cost-utility" as a measure of the efficiency of a test. J. Amer. Statist. Ass., 1947, 42, 246-255.
6. Blackwell, David, and Girshick, M. A. Theory of games and statistical decisions. New York: Wiley, 1954.
7. Bloom, B. S. Test reliability for what? J. educ. Psychol., 1942, 33, 517-526.
8. Bordin, E. S., and Bixler, R. H. Test selection: a process of counseling. Educ. psychol. Measmt, 1946, 6, 361-373.
9. Box, G. E. P. The exploration and exploitation of response surfaces: some general considerations and examples. Biometrics, 1954, 10, 16-60.
10. Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. J. educ. Psychol., 1946, 37, 65-76.
11. Brogden, H. E. When testing pays off. Personnel Psychol., 1949, 2, 171-185.
12. Brogden, H. E. A new coefficient: application to biserial correlation and to estimation of selective efficiency. Psychometrika, 1949, 14, 169-182.
13. Brogden, H. E. Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educ. psychol. Measmt, 1951, 11, 173-196.
14. Brogden, H. E., and Taylor, E. K. The dollar criterion -- applying the cost accounting concept to criterion construction. Personnel Psychol., 1950, 3, 133-154.
15. Bross, I. D. J. Design for decision. New York: Macmillan, 1953.

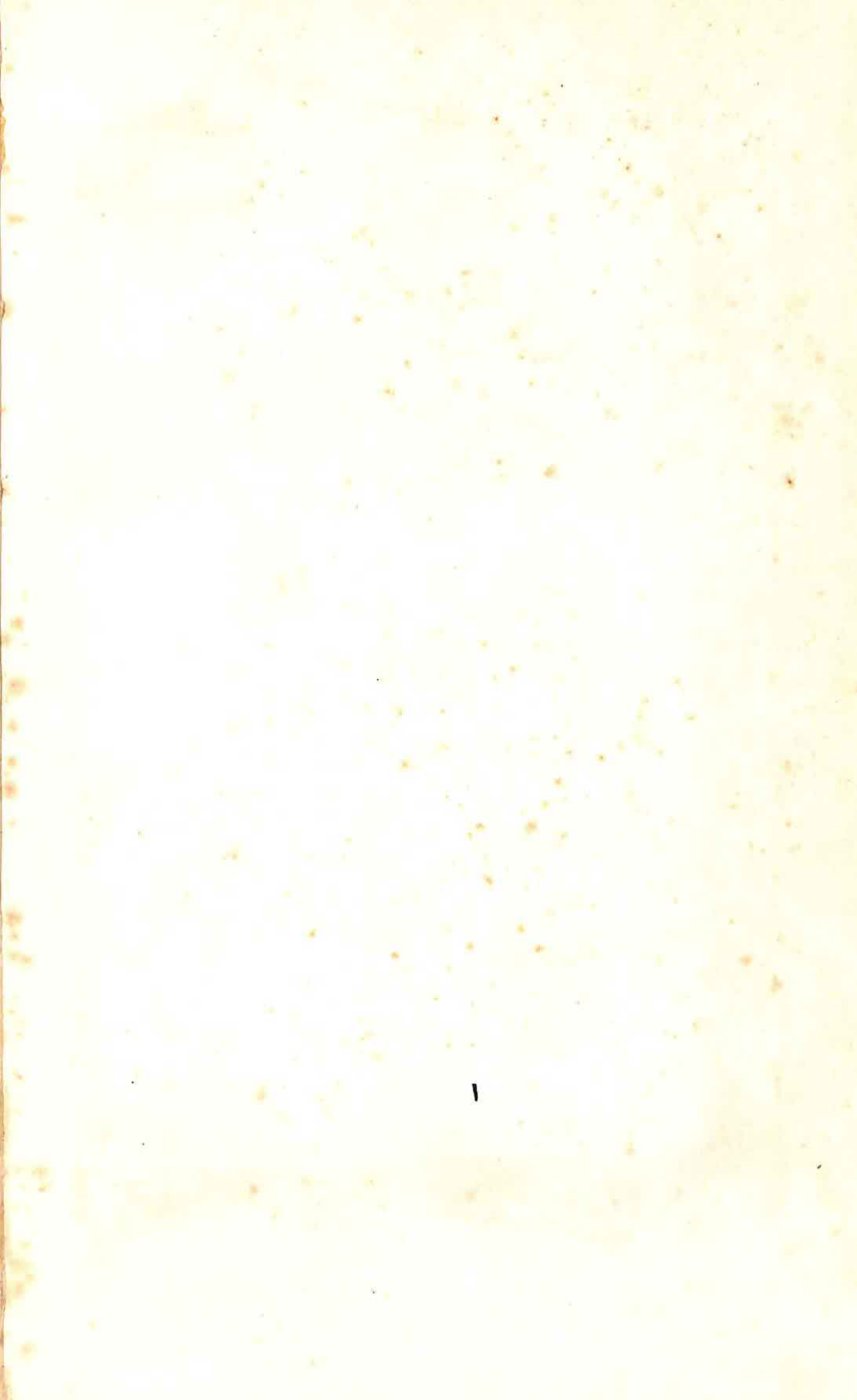
16. Brown, C. W., and Ghiselli, E. E. Per cent increase in proficiency resulting from use of selective devices. J. appl. Psychol., 1953, 37, 341-345.
17. Cardinet, Jean. Problèmes relatifs à la repartition du personnel. Paris, the author, 1955. Mimeographed.
18. Chernoff, Herman. Locally optimal designs for estimating parameters. Ann. math. Stat., 1953, 24, 586-602.
19. Churchman, C. W., and Ackoff, R. L. An approximate measure of value. J. Operations Res. Soc. Amer., 1954, 2, 172-187.
20. Cochran, W. G. Improvement by means of selection. In J. Neyman (Ed.), Second Berkeley symposium on mathematical statistics and probability. Berkeley, Calif.: University of California Press, 1951. Pp. 449-470.
21. Conrad, Herbert. Information which should be provided by test publishers and testing agencies on the validity and use of their tests. In Proceedings, 1949 invitational conference on testing problems. Princeton, N. J.: Educational Testing Service, 1950. Pp. 63-68.
22. Coombs, C. H. Theory and methods of social measurement. In L. Festinger and D. Katz (Eds.), Research methods in the behavioral sciences. New York: Dryden Press, 1953. Pp. 471-563.
23. Cronbach, L. J. A consideration of information theory and utility theory as tools for psychometric problems. Urbana, Ill.: Coll. of Educ., University of Illinois, 1953.
24. Cronbach, L. J. On the non-rational application of information measure in psychology. In H. Quastler (Ed.), Information theory in psychology: problems and methods. Glencoe, Ill.: Free Press, 1956. Pp. 14-30.
25. Cronbach, L. J. The counselor's problems from the perspective of communication theory. In Vivian H. Hewer (Ed.), New perspectives in counseling. Minneapolis: University of Minnesota Press, 1955. Pp. 3-19.
26. Cronbach, L. J. Processes affecting scores on "understanding of others" and "assumed similarity". Psychol. Bull., 1955, 52, 177-193.
27. Doppelt, J. E., and Bennett, G. K. Reducing the cost of training satisfactory workers by using tests. Personnel Psychol., 1953, 6, 1-9.
28. Duncan, O. D., Ohlin, L. E., Reiss, A. J., Jr., and Stanton, H. R. Formal devices for making selection decisions. Amer. J. Sociol., 1953, 58, 573-585.
29. Dwyer, P. S. Selection and linear combination of tests in relation to multiple criteria and differential classification. PRB Res. Note 7, Department of the Army, 1953.
30. Dwyer, P. S. Multiple assignments of persons to jobs. In Proceedings, 1954 invitational conference on testing problems. Princeton, N. J.: Educational Testing Service, 1955. Pp. 11-29.
31. Edwards, Ward. The theory of decision making. Psychol. Bull., 1954, 51, 380-418.
32. Elfving, Gustav. Optimum allocation in linear regression. Ann. math. Stat., 1952, 23, 255-263.

33. Evans, R. N. A suggested use of sequential analysis in performance acceptance testing. Urbana, Ill.: Coll. of Educ., University of Illinois, [1953].
34. Fiske, D. W., and Jones, L. V. Sequential analysis in psychological research. Psychol. Bull., 1954, 51, 264-276.
35. Freeman, H. A. and others (Eds.). Sampling inspection. New York: McGraw-Hill, 1948.
36. Girshick, M. A. An elementary survey of statistical decision theory. Rev. educ. Res., 1954, 24, 448-466.
37. Goodman, L. A. The use and validity of a prediction instrument. I. A reformulation of the use of a prediction instrument. Amer. J. Sociol., 1953, 58, 503-508.
38. Goodman, L. A. The use and validity of a prediction instrument. II. The validation of prediction. Amer. J. Sociol., 1953, 58, 510-513.
39. Horst, Paul. Determination of optimal test length to maximize the multiple correlation. Psychometrika, 1949, 14, 79-88.
40. Horst, Paul. A technique for the development of a differential prediction battery. Psychol. Monogr., 1954, 68, No. 9 (Whole No. 380).
41. Horst, Paul. Optimal test length for maximum differential prediction. Psychometrika, 1956, 21, 51-67.
42. Horst, Paul, and MacEwan, Charlotte. Optimal test length for maximum absolute prediction. Psychometrika, 1956, 21, 111-125.
43. Hull, C. L. Aptitude testing. Yonkers-on-Hudson: World Book, 1928.
44. Hunt, W. A., Wittson, C. L., and Hunt, E. B. Military performance of a group of marginal neuropsychiatric cases. Amer. J. Psychiat., 1952, 109, 168-171.
45. Jarrett, R. F. Per cent increase in output of selected personnel as an index of test efficiency. J. appl. Psychol., 1948, 32, 135-146.
46. Kelley, T. L. Statistical method. New York: Macmillan, 1923.
47. Koopman, B. O. The optimum distribution of effort. J. Operations Res. Soc. Amer., 1953, 1, 52-63.
48. Loevinger, Jane, Gleser, Goldine C., and DuBois, P. H. Maximizing the discriminating power of a multiple-score test. Psychometrika, 1953, 18, 309-317.
49. Long, W. F., and Burr, I. W. Development of a method for increasing the utility of multiple correlations by considering both testing time and test validity. Psychometrika, 1949, 14, 137-161.
50. Lord, F. M. Some perspectives on "the attenuation paradox in test theory". Psychol. Bull., 1955, 52, 505-511.
51. Lubin, Ardie. Linear and non-linear discriminating functions. Brit. J. Psychol., 1950, 3, 90-104.
52. Magwire, C. A. Sequential decisions involving the choice of experiments. Stanford, Calif.: Stanford University, 1953. Unpublished technical report.
53. Meehl, P. E. Clinical vs. statistical prediction. Minneapolis: University of Minnesota Press, 1954.

54. Raines, G.N., Wittson, C. L., Hunt, W. A., and Herrmann, R. S. Psychiatric selection for military service. J. Amer. med. Ass., 1954, 156, 817-821.
55. Richardson, M. W. The interpretation of a test validity coefficient in terms of increased efficiency of a selected group of personnel. Psychometrika, 1944, 9, 245-248.
56. Sarbin, T. R. A contribution to the study of actuarial and individual methods of prediction. Amer. J. Sociol., 1942, 48, 593-602.
57. Shannon, C. E., and Weaver, Warren. The mathematical theory of communication. Urbana, Ill.: University of Illinois Press, 1949.
58. Smith, N. M., Jr., Walters, S. S., Brooks, F. C., and Blackwell, D. H. The theory of value and the science of decision, a summary. J. Operations Res. Soc. Amer., 1953, 1, 103-113.
59. Sobel, Milton, and Wald, Abraham. A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. Ann. math. Stat., 1949, 20, 502-522.
60. Somerville, P. N. Some problems of optimum sampling. Biometrika, 1954, 41, 420-429.
61. Statistical Research Group, Columbia University, Sequential analysis of statistical data, applications. New York: Columbia University Press, 1945.
62. Stein, Charles, and Wald, Abraham. Sequential confidence intervals for the mean of a normal distribution with known variance. Ann. math. Stat., 1947, 18, 427-433.
63. Taylor, C. W. Maximizing predictive efficiency for a fixed total testing time. Psychometrika, 1950, 15, 391-406.
64. Taylor, H. C., and Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection. J. appl. Psychol., 1939, 23, 565-578.
65. Thrall, R. M., Coombs, C. H., and Davis, R. L. (Eds.). Decision processes. New York: Wiley, 1954.
66. Thurstone, L. L. Reliability and validity of tests. Ann Arbor, Mich.: Edwards Brothers, 1931.
67. von Neumann, John, and Morgenstern, Oskar. Theory of games and economic behavior. Princeton, N.J.: Princeton University Press, 1947.
68. Votaw, D. F., Jr., and Dailey, J. T. Assignment of personnel to jobs. Human Resources Res. Center, Res. Bull., 52-24, 1952.
69. Wald, Abraham. Statistical decision functions. New York: Wiley, 1950.
70. Wesman, A. G. Better than chance. Test Service Bull., 1953, No. 45, 1-5.
71. Wesman, A. G., and Bennett, G. K. Problems of differential prediction. Educ. psychol. Measmt, 1951, 11, 265-273.
72. Wolfowitz, Jacob. Minimax estimates of the mean of a normal distribution with known variance. Ann. math. Stat., 1950, 21, 218-230.







Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological
Research Library.**

The book is to be returned within
the date stamped last.

30 JUN 1934

WBGP-59/60-5119C-5M

151.2

CRO

Form No. 4

BOOK CARD

Coll. No. 151.2/CRO Accn. No. 1327

Author. ~~Port~~ Cronbach

Title... Psychol test & Personnel
Decisions.

Date.

Issued to

Returned on

30 JUN 1951

151.2

CRO

